Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Matematica

PCA E CLUSTERING APPLICATI A UN DATASET SUL NUOTO

Tesi di Laurea in Matematica Computazionale

Relatore:
Chiar.mo Prof.
VALERIA SIMONCINI

Presentata da:
ALICE SENATORE

V Sessione
Anno Accademico 2024-2025

Indice

Introduzione

Negli ultimi decenni, la disponibilità sempre crescente di dati, favorita dallo sviluppo di sistemi automatici di raccolta e archiviazione, ha reso indispensabile lo sviluppo e l'impiego di tecniche statistiche e numeriche in grado di estrarre informazione utile da insiemi di dati sempre più vasti e complessi.
In questo contesto si inserisce il Data Mining, disciplina che analizza grandi moli di dati, per estrarne informazione.
Con il termine 'informazione' si intende il risultato di un processo di trasformazione, interpretazione ed estrazione dei dati, dove questi ultimi, originariamente grezzi e privi di un significato immediatamente evidente, vengono elaborati per rivelare pattern, tendenze o correlazioni.
L'analisi di dati e la ricerca di modelli o regolarità è sempre stata effettuata; ma è solo negli ultimi anni che sono stati sviluppati strumenti computazionali capaci di esplorare strutture anche molto complesse.
Tra le strategie più diffuse di analisi di dati rientrano:

- *(Caratterizzazione)* Si determinano proprietà comuni di gruppi di dati;
- *(Discriminanza)* Si confrontano caratteristiche diverse tra gruppi di dati;
- *(Appartenenza)* Si riconoscono nuovi dati come membri di determinati gruppi.

In numerosi campi applicativi, dalla biologia alla finanza, dalle scienze sociali allo sport, l'analisi multivariata e i metodi di riduzione dimensionale rappresentano ormai strumenti fondamentali per interpretare fenomeni caratterizzati da molte variabili tra loro correlate.

Tra le tecniche maggiormente utilizzate rivestono un ruolo centrale l'Analisi delle Componenti Principali, che permette di sintetizzare la variabilità del dataset riducendo il numero di variabili mediante la costruzione di nuove direzioni dominanti, e le tecniche di clustering, che permettono di riconoscere similarità o dissimilarità tra osservazioni o variabili e che mirano a individuare gruppi omogenei all'interno e disomogenei all'esterno.
La finalità di questa tesi è duplice.
Da un lato presentare teoricamente gli strumenti utilizzati: nel secondo capitolo si introduce l'algoritmo di PCA, le metodologie di scelta delle componenti principali e la loro interpretazione grafica; nel terzo capitolo si analizzano le tecniche di clustering, concentrandosi in particolare sul Complete Linkage per i metodi gerarchici e sulle K-medie per quelli non gerarchici, con relative interpretazioni grafiche.
Dall'altro lato applicare questi strumenti: nel quarto capitolo viene fatta un'analisi su un dataset reale relativo al nuoto agonistico, fornito da un dottorando del Dipartimento di Scienze per la Qualità della Vita. Il dataset analizzato comprende un ampio insieme di variabili antropometriche e prestative, rilevate su nuotatori e nuotatrici di età, proporzioni fisiche e stili di nuoto più efficaci differenti e, dal momento in cui non tutti i soggetti presentano dati completi, l'analisi viene svolta in due fasi: su un dataset ristretto, ottenuto selezionando solo gli individui con tutte le variabili nello stile libero disponibili; sul dataset completo, affrontando il problema dei dati mancanti con metodi adeguati.
L'obiettivo complessivo è quindi quello di mostrare come tecniche dell'analisi numerica, della statistica multivariata e del data mining possano essere efficacemente utilizzate per comprendere fenomeni complessi, rivelare strutture interne ai dati e supportare interpretazioni coerenti in un contesto reale e applicativo, come quello del nuoto agonistico.

Capitolo 1

Nozioni Preliminari

1.1 Le matrici

Definizione 1.1.1. Dato un campo $\mathbb{R}$, si definisce lo spazio dei vettori colonna come l'insieme i cui elementi sono successioni in colonna di n numeri di $\mathbb{R}$:

$$\mathbb{R}^n = \left\{ \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \;\middle|\; a_1, a_2, \ldots, a_n \in \mathbb{R} \right\}.$$

Definizione 1.1.2. Presi m, n, numeri interi positivi, una matrice $m \times n$ a coefficienti in $\mathbb{R}$ è un insieme di mn elementi di $\mathbb{R}$ disposti in questo modo:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

dove m rappresenta il numero di righe e n il numero di colonne.

In questa tesi considereremo come campo quello dei numeri reali $\mathbb{R}$ e denoteremo quindi con $\mathbb{R}^{m\times n}$ lo spazio delle matrici a coefficienti reali con m righe e n colonne e con $\mathbb{R}^m$ lo spazio dei vettori colonna a coefficienti reali.

Definizione 1.1.3. La trasposizione di un vettore $v \in \mathbb{R}^m$ è l'operazione di scambio delle sue righe con le sue colonne:

$$\begin{pmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{m1} \end{pmatrix}^T = \begin{pmatrix} v_{11} & v_{21} & \cdots & v_{m1} \end{pmatrix}.$$

Definizione 1.1.4. Una matrice $A \in \mathbb{R}^{n\times n}$ si dice diagonale se è nulla al di fuori della diagonale principale: $a_{ij} = 0 \quad \forall i \neq j$.

Definizione 1.1.5. La matrice identità I è la matrice diagonale con tutti i coefficienti sulla diagonale principale uguali a 1.

Definizione 1.1.6. Data $A \in \mathbb{R}^{n\times n}$ la traccia di A è la somma dei suoi elementi diagonali: $tr(A) = \sum_{i=1}^{n} a_{ii}$.

Definizione 1.1.7. Una matrice $A \in \mathbb{R}^{n\times n}$ si dice invertibile o non singolare se esiste $A^{-1} \in M_n(\mathbb{K})$ tale che $A^{-1}A = AA^{-1} = I$.

Osservazione 1.1.1. Se $A \in M_n(\mathbb{K})$ è invertibile, allora anche A^T è invertibile con inversa $(A^T)^{-1} = (A^{-1})^T$.

Definizione 1.1.8. Una matrice $A \in \mathbb{R}^{n\times n}$ si dice simmetrica se $A = A^T$.

Definizione 1.1.9. Data una matrice $A \in \mathbb{R}^{n\times n}$la forma bilineare associata ad A è l'applicazione $\beta_A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ tale che $\beta_A(x, y) = x^T AY, \quad x, y \in \mathbb{R}^n$.

Definizione 1.1.10. Una matrice $A \in \mathbb{R}^{n\times n}$ si dice semidefinita positiva se $x^T Ax \geq 0 \quad \forall x \in \mathbb{R}^n$; si dice definita positiva se $x^T Ax > 0 \quad \forall x \in \mathbb{R}^n \setminus \{0\}$.

Definizione 1.1.11. Una matrice quadrata $A \in \mathbb{R}^{n\times n}$ si dice ortogonale se $A^T A = AA^T = I$, cioè se A è invertibile e $A^T = A^{-1}$.

Definizione 1.1.12. Una funzione $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}$ è una norma vettoriale se, per ogni $x, y \in \mathbb{R}^n$, soddisfa le seguenti proprietà:

1. $\|x\| \geq 0$ e $\|x\| = 0 \iff x = 0$, dove 0 indica il vettore nullo;
2. $\|\alpha x\| = |\alpha| \cdot \|x\| \quad \forall \alpha \in \mathbb{R}$;
3. *(disuguaglianza triangolare)* $\|x + y\| \leq \|x\| + \|y\|$.

Definizione 1.1.13. La norma-p vettoriale è definita come:

$$\|x\|_p := (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}} \quad \forall x \in \mathbb{R}^n; p \geq 1.$$

In particolare si hanno le seguenti norme:

- *(Norma-1 o del modulo)* $\|x\|_1 = \sum_{i=1}^{n} |x_i| \quad \forall x \in \mathbb{R}^n$;
- *(Norma-2 o euclidea)* $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \quad \forall x \in \mathbb{R}^n$;
- *(Norma-∞ o del massimo)* $\|x\|_\infty = \max_i |x_i| \quad \forall x \in \mathbb{R}^n$.

Definizione 1.1.14. Una funzione $\|\cdot\| \colon \mathbb{R}^{n \times n} \to \mathbb{R}$ è una norma di matrice se, per ogni $A, B \in \mathbb{R}^{n \times n}$, soddisfa le seguenti proprietà:

1. $\|A\| \geq 0$ e $\|A\| = 0 \iff A = 0$, dove 0 indica la matrice nulla;
2. $\|\alpha A\| = |\alpha| \cdot \|A\| \quad \forall \alpha \in \mathbb{R}$;
3. $\|A + B\| \leq \|A\| + \|B\|$;
4. $\|AB\| \leq \|A\| \, \|B\|$, da cui $\|A^m\| \leq \|A\|^m, \forall m \in \mathbb{N}$.

La definizione può essere generalizzata al caso rettangolare, con le opportune modifiche sulle dimensioni delle matrici.

Definizione 1.1.15. La norma-p matriciale indotta dalla norma.p vettoriale è definita come:

$$\|A\|_p := \max_{0 \neq x \in \mathbb{R}^n} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p = 1} \|Ax\|_p \quad \forall A \in \mathbb{R}^{n \times n}.$$

Definizione 1.1.16. La norma di Frobenius è definita come:

$$\|A\|_F := \left(\sum_{i,j} |a_{ij}|^2\right)^{\frac{1}{2}} \quad \forall A \in \mathbb{R}^{n\times n}.$$

Proposizione 1.1.1. *Alcune proprietà delle norme di matrice sono:*

1. $\|A\| = \|A^T\|, \quad \forall A \in \mathbb{R}^{n\times n}$;
2. *Se* $\|\cdot\|$ *è una norma matriciale indotta, allora* $\|Ax\| \leq \|A\|\,\|x\|, \quad \forall x \in \mathbb{R}^n$;
3. *Se* $\|\cdot\|$ *è una norma matriciale indotta, allora* $\|I\| = 1$ *dove* I *indica la matrice identità; e per una qualsiasi norma matriciale vale* $\|I\| \geq 1$;
4. $\|A^{-1}\| \geq \dfrac{1}{\|A\|}, \quad \forall A \in \mathbb{R}^{n\times n}$ *matrice invertibile;*
5. $\|A\|_F = tr(A^*A)^{\frac{1}{2}}, \quad \forall A \in \mathbb{R}^{n\times n}$;
6. *Per ogni* $U, V \in \mathbb{C}^{n\times n}$ *unitarie* $\|UAV^*\|_F = \|A\|_F, \quad A \in \mathbb{R}^{n\times n}$;
7. *Se* $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}$, *allora* $\|D\|_F = (\sum_{i=1}^{n} \sigma_i^2)^{\frac{1}{2}}$;
8. *Se* $A \in \mathbb{R}^{n\times n}$ *è non singolare allora, per ogni norma indotta, si ha che* $\min_{\|x\|=1} \|Ax\| = \dfrac{1}{\|A^{-1}\|}$.

1.2 Il problema agli autovalori

Definizione 1.2.1. Sia $A \in \mathbb{R}^{m\times n}$, un autovettore è un vettore non nullo $x \in \mathbb{C}^n$ tale che $Ax = \lambda x$ con $\lambda \in \mathbb{C}$ autovalore di A. Inoltre, la coppia (λ, x) viene detta autocoppia di A.

Definizione 1.2.2. Sia $A \in \mathbb{R}^{n\times n}$, si definisce spettro di A l'insieme:

$$spec(A) = \Lambda(A) := \{\lambda \in \mathbb{C} \mid \lambda \text{ autovalore di } A\}.$$

Definizione 1.2.3. Sia $A \in \mathbb{R}^{n\times n}$, si definisce raggio spettrale di A:

$$\rho(A) := \max\{|\lambda|| \ \lambda \in \Lambda(A)\}$$

e misura la massima distanza di $\Lambda(A)$ dall'origine.

Definizione 1.2.4. Sia $A \in \mathbb{R}^{n\times n}$ simmetrica, si definisce quoziente di Rayleight il rapporto $\dfrac{x^TAx}{x^Tx}$; $\quad x \in \mathbb{R}^n \setminus \{0\}$.
In questo caso il quoziente di Rayleight è reale.

Definizione 1.2.5. Una matrice $A \in \mathbb{R}^{n\times n}$ si dice diagonalizzabile se è simile a una matrice diagonale, ovvero se esiste una matrice $P \in \mathbb{R}^{n\times n}$) invertibile tale che $PD = AP$; $\quad D \in \mathbb{R}^{n\times n}$ diagonale.

Definizione 1.2.6. Una matrice $A \in \mathbb{R}^{n\times n}$ si dice normale se soddisfa:

$$A^TA = AA^T.$$

Proposizione 1.2.1. *Per una matrice $A \in \mathbb{R}^{n\times n}$ valgono le seguenti decomposizioni:*

- *Se A è simmetrica, $A = Q\Lambda Q^T$ con $Q \in \mathbb{R}^{n\times n}$ ortogonale, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$ dove $\lambda_1, \ldots, \lambda_n$ sono gli autovalori di A;*
- *Se A è normale, $A = QDQ^T$ con $Q \in \mathbb{R}^{n\times n}$ ortogonale, $D \in \mathbb{R}^{n\times n}$ diagonale;*
- *Se A non è diagonalizzabile, è sempre possibile scrivere la decomposizione di Schur: $A = QRQ^T$ con $Q \in \mathbb{R}^{n\times n}$ ortogonale e R triangolare superiore avente sulla diagonale gli autovalori di A.*

Data la matrice $A \in \mathbb{R}^{n\times n}$, il problema degli autovalori consiste nel calcolo della coppia (λ, x) con $\lambda \in \mathbb{C}$ e $0 \neq x \in \mathbb{R}^n$ tale che $Ax = \lambda x$.
Questo problema è non lineare, quindi non ci sono in generale algoritmi "diretti" per la sua risoluzione.

Osservazione 1.2.1. Dopo l'introduzione della definizione di autovalore è possibile riscrivere alcune caratterizzazioni:

- Una matrice $A \in \mathbb{R}^{n\times n}$ è non invertibile o singolare se e solo se $0 \in \Lambda(A)$;
- Se $A \in \mathbb{R}^{n\times n}$ è diagonale, i suoi elementi diagonali coincidono con i suoi autovalori e gli autovettori relativi sono la base canonica;
- Se $\lambda \in \Lambda(A)$, allora $\lambda^{-1} \in \Lambda(A^{-1})$, ma l'autovettore relativo rimane lo stesso;
- Sia $A \in \mathbb{R}^{n\times n}$ simmetrica, allora A ha solo autovalori reali;
- Il determinante di $A \in \mathbb{R}^{n\times n}$ soddisfa $det(A) = \prod_{i=1}^{n} \lambda_i$;
- La traccia di $A \in \mathbb{R}^{n\times n}$ soddisfa $tr(A) = \sum_{i=1}^{n} \lambda_i$;
- Se $A \in \mathbb{R}^{n\times n}$ è simmetrica e definita positiva, allora $\lambda_i > 0 \quad \forall i$. Se è semidefinita positiva $\lambda_i \geq 0 \quad \forall i$ (ugualmente se definita negativa e semidefinita negativa).

Teorema 1.2.1 (Rayleight-Ritz)**.** *Sia $A \in \mathbb{R}^{n\times n}$ simmetrica, $\lambda_1 \leq \ldots \leq \lambda_n$ autovalori di A, vale:*

$$\lambda_1 \leq \frac{v^T A v}{v^T v} \leq \lambda_n \qquad \forall 0 \neq v \in \mathbb{R}^n.$$

Inoltre i valori di λ_1 e λ_n vengono effettivamente raggiunti rispettivamente per $v = x_{min}$ e $v = x_{max}$, ovvero l'autovettore più piccolo e quello più grande:

$$\lambda_{max} = \lambda_n = \max_{x\neq 0} \frac{x^T A x}{x^T x}, \quad \lambda_{min} = \lambda_1 = \min_{x\neq 0} \frac{x^T A x}{x^T x}.$$

Dimostrazione. A simmetrica si decompone: $A = X\Lambda X^T$ con $X = [x_1, \ldots, x_n] \in \mathbb{R}^{n\times n}$ ortogonale e $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$.

Sia $0 \neq v \in \mathbb{R}^n$, vale $\dfrac{v^T A v}{v^T v} = \dfrac{v^T X \Lambda X^T v}{v^T X X^T v} = \dfrac{w^T \Lambda w}{w^T w} = \dfrac{\sum_{i=1}^{n} \lambda_i |w_i|^2}{\|w\|^2}$.
Essendo $\|w\| = \|v\|$ perchè X ortogonale,
si ha $\sum_{i=1}^{n} \lambda_i |w_i|^2 \leq \lambda_{max} \sum_{i=1}^{n} |w_i|^2 \leq \lambda_{max} \|w\|^2 = \lambda_{max} \|v\|^2$.

Avendo ordinato gli autovalori in ordine decrescente, allora $\lambda_{max} = \lambda_n$
e si può concludere: $\frac{v^T A v}{v^T v} \leq \lambda_n \qquad \forall 0 \neq v \in \mathbb{R}^n$.
Conti analoghi permettono di ottenere l'altra disuguaglianza. □

Corollario 1.2.1. *Siano* $A = X\Lambda X^T$, $\lambda_1 \leq \ldots \leq \lambda_n$ *autovalori di* A, $X = [x_1, \ldots, x_n] \in \mathbb{R}^{n\times n}$ *ortogonale,* $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$ *e sia* (λ_1, x_1) *la più piccola autocoppia di* A*; allora* $\min\limits_{0\neq x\perp x_1} \frac{x^T A x}{x^T x} = \lambda_2$.

Dimostrazione. Sia $x \perp x_1 \Rightarrow x_1^T x = 0$ e, definendo $q := X^T x$, si ha $\|x\| = \|q\|$.
Quindi $x^T A x = \sum\limits_{i=1}^{n} \lambda_i |q_i^T q_i|^2 = \sum\limits_{i>1} \lambda_i |q_i^T q_i|^2 \geq \lambda_2 \sum\limits_{i>1} |q_i^T q_i|^2 = \lambda_2 \sum\limits_{i=1}^{n} |q_i^T q_i|^2 =$
$\lambda_2 x^T x$
L'inf di $x^T A x$ viene raggiunto e si ha il min. □

Osservazione 1.2.2. Il risultato del corollario si può generalizzare:

$$\min_{0\neq x\perp x_1,\ldots,x_{k-1}} \frac{x^T A x}{x^T x} = \lambda_k; \quad k = 2, \ldots, n.$$

Capitolo 2

Analisi delle componenti principali

2.1 Elementi di statistica descrittiva multivariata

In genere, con il termine statistica si intende la disciplina che studia le tecniche per la raccolta dei dati e la loro elaborazione, in modo da ottenere il più elevato numero di informazioni in riferimento al fenomeno in studio. Quando si raccolgono informazioni in riferimento ad un certo fenomeno, ci si trova ad avere a che fare con una mole notevole di dati grezzi, di conseguenza, il primo problema che ci si trova ad affrontare è quello di sintetizzare la massa di dati grezzi in pochi numeri o indicatori particolarmente informativi, utilizzando metodiche grafiche o numeriche che siano in grado di descrivere la massa di dati senza alterarne il senso complessivo. Questa parte della statistica è nota con il nome di statistica descrittiva.

Definizione 2.1.1. Si dice matrice dei dati la matrice $X \in \mathbb{R}^{n\times p}$, dove n sono le osservazioni e p le variabili.

Definizione 2.1.2. Sia $X \in \mathbb{R}^{n\times p}$ matrice dei dati, si dice media campionaria il vettore riga $\bar{x} = [\bar{x}_1, \ldots, \bar{x}_p] \in \mathbb{R}^{1\times p}$ dove $\bar{x}_i = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} x_{ji} \quad \forall i = 1, \ldots, p$ è la media campionaria degli elementi della i-esima colonna di X, cioè dell'i-esima variabile.

Definizione 2.1.3. Sia $X \in \mathbb{R}^{n\times p}$ matrice dei dati, si dice matrice di covarianza campionaria la matrice $S \in \mathbb{R}^{p\times p}$:

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{pmatrix}, \quad s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \quad 1 \leq i,k \leq p.$$

Osservazione 2.1.1. La matrice di covarianza campionaria è un indice di dispersione, ovvero una misura statistica che descrive quanto i dati in un campione sono sparsi attorno a un valore centrale: la media.
Il coefficiente s_{ik} è grande se entrambe x_{ji}, x_{jk} sono dispersive.

Definizione 2.1.4. Sia $S \in \mathbb{R}^{p\times p}$ la matrice di covarianza campionaria, la varianza totale campionaria è data da: $tr(S) := \sum_{i=1}^{p} s_{ii}$.

Osservazione 2.1.2. La varianza totale campionaria non è molto indicativa perchè non tiene conto della covarianza tra variabili.

Definizione 2.1.5. Sia $X \in \mathbb{R}^{n\times p}$ matrice dei dati, si dice matrice di correlazione campionaria la matrice $R \in \mathbb{R}^{p\times p}$:

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{12} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & r_{pp} \end{pmatrix}, \quad r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}s_{kk}}} \quad 1 \leq i,k \leq p.$$

Osservazione 2.1.3. La matrice di correlazione campionaria è una matrice simmetrica semidefinita positiva e adimensionale che coglie solo la variabilità lineare dei dati.
La correlazione tra due variabili è un numero appartenente all'intervallo $[-1, 1]$ ed è positiva quando al crescere di una variabile l'altra cresce; negativa quando al crescere di una variabile l'altra cala.

Definizione 2.1.6. Sia $X \in \mathbb{R}^{n\times p}$ matrice dei dati, $R \in \mathbb{R}^{p\times p}$ matrice di correlazione campionaria e $\{(\lambda_i, v_i)\}_{i=1,\dots,p}$ autocoppie di R; il rapporto di variabilità è definito come: $rv \in \mathbb{R}^p$ tale che $rv_i = \dfrac{1}{tr(R)} \displaystyle\sum_{k\leq i} \lambda_k$ e rappresenta l'importanza dei primi i autovalori rispetto a tutti.

2.2 Componenti principali di un campione di dati

La Principal Component Analysis (PCA) è una tecnica di apprendimento non supervisionato utilizzata nella moderna analisi di dati e riguardante diversi campi di ricerca: dalle neuroscienze alla computer graphic.
È un metodo relativamente semplice per l'estrazione di informazioni rilevanti da dati di difficile interpretazione; l'idea alla base della PCA è di ridurre la dimensionalità del dataset, mantenendo quanta più varianza possibile nei dati. La riduzione viene fatta passando da un set di variabili di dimensione $n \times p$ a un nuovo set di variabili di dimensione $n \times k$ con $k << p$ latenti (ovvero non misurabili), non correlate tra loro e ordinate in modo che le prime mantengano la maggior parte della varianza presente in tutte le variabili originali.
In altre parole, l'obiettivo della PCA è trovare una base vettoriale alternativa $\{y_1, \dots, y_p\}$, combinazione lineare della base originale $\{x_1, \dots, x_p\}$, che meglio esprima le proprietà del data set, filtrando il rumore e rivelando la struttura prima dei dati.

Sia $X = [x_1, \dots, x_p] \in \mathbb{R}^{n\times p}$ matrice dei dati, $S \in \mathbb{R}^{p\times p}$ matrice di covarianza, $\lambda_1 \geq \dots \geq \lambda_p \geq 0$ autovalori della matrice di covarianza e $R \in \mathbb{R}^{p\times p}$ matrice di correlazione, lo scopo è determinare $A = [a_1, \dots, a_p] \in \mathbb{R}^{n\times p}$ matrice dei coefficienti tale che $Y = XA$ con $Y = [y_1, \dots, y_p] \in \mathbb{R}^{n\times p}$, con $\{y_1, \dots, y_p\}$ non correlate (covarianza nulla) e che massimizzino la variabilità che ognuna delle p variabili ha nel campione considerato.

Le p colonne di Y sono combinazione lineare delle p colonne di X:

$$\begin{cases} y_1 = Xa_1 = x_1a_{11} + x_2a_{12} + \ldots + x_pa_{1p} \\ y_2 = Xa_2 = x_1a_{21} + x_2a_{22} + \ldots + x_pa_{2p} \\ \vdots \\ y_p = Xa_1 = x_1a_{p1} + x_2a_{p2} + \ldots + x_pa_{pp} \end{cases}$$

con varianza campionaria $var(y_i) = a_i^T S a_i \quad \forall i = 1, \ldots, p$ e covarianza campionaria $cov(y_i, y_j) = a_i^T S a_j \quad \forall i, j = 1, \ldots, p, i \neq j$.
La prima componente principale è $y_1 = Xa_1$, che massimizza $var(y_1) = a_1^T S a_1$.
La seconda componente principale è $y_2 = Xa_2$, che massimizza $var(y_2) = a_2^T S a_2$ e risulta ortogonale in senso di covarianza a y_1, ossia soddisfa $cov(y_1, y_2) = a_1^T S a_2 = 0$.
E, in generale, la i-esima componente principale è $y_i = Xa_i$, che massimizza $var(y_i) = a_i^T S a_i$ e risulta ortogonale in senso di covarianza a y_k con $k < i$, ossia soddisfa $cov(y_k, y_i) = a_k^T S a_i = 0 \quad \forall k < i$.
Trovare a_1 che massimizzi la varianza campionaria, significa trovare:

$$\max_{a_1 \in \mathbb{R}, \|a_1\|=1} a_1^T S a_1.$$

Per il Teorema 1.2.1 tale massimo è il più grande autovalore λ_1 della matrice di covarianza, ottenuto scegliendo a_1 come primo autovettore di S.
Trovare a_2 che massimizzi la varianza campionaria, sotto il vincolo che la seconda componente principale y_2 sia non correlata a y_1, significa trovare:

$$\max_{a_2 \in \mathbb{R}, \|a_2\|=1} a_2^T S a_2 \mid a_1^T S a_2 = 0$$

cioè, unendo queste due condizioni, trovare:

$$\max_{a_2 \in \mathbb{R}, \|a_2\|=1, Xa_2 \perp Xa_1} a_2^T S a_2.$$

Per il Corollario 1.2.1 si è osservato che tale massimo è il più grande autovalore escluso λ_1 (ovvero λ_2) della matrice di covarianza, ottenuto scegliendo a_2 come secondo autovettore di S.
Iterando questo ragionamento, si determina a_p, osservando che $A = V$ con $V =$

$[v_1, \ldots, v_p]$ matrice degli autovettori della matrice di covarianza relativi agli autovalori $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$.

Proposizione 2.2.1. *Data $S \in \mathbb{R}^{p\times p}$ matrice di covarianza e $(\lambda_1, v_1), \ldots, (\lambda_p, v_p)$ autocoppie di S tali che $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$; definendo le componenti principali $y_i = Xa_i = Xv_i$ come sopra, allora:*

1. *(Varianza campionaria) $var(y_i) = \lambda_i \quad \forall i = 1, \ldots, p$;*
2. *(Covarianza campionaria) $cov(y_k, y_i) = 0 \quad \forall i \neq k$;*
3. *(Varianza totale campionaria) $\sum_{i=1}^{p} \lambda_i$.*

Dimostrazione. Sia $y_i = Xa_i$ con $a_i \in \mathbb{R}^p, \|a_i\| = 1$, ricordando che la i-esima componente principale è stata ottenuta scegliendo $a_i = v_i$ con v_i autovettore di S,

1. $var(y_i) = a_i^T S a_i = v_i^T S v_i = v_i^T(\lambda_i v_i) = \lambda_i(v_i^T v_i) = \lambda_i$;
2. $cov(y_k, y_i) = a_k^T S a_i = v_k^T S v_i = v_k^T(\lambda_i v_i) = \lambda_i(v_k^T v_i) = 0$, perchè gli autovettori di S formano una base ortonormale;
3. per il sesto punto dell'Osservazione 1.2.1, la traccia della matrice S è pari alla somma dei suoi autovalori.

□

Osservazione 2.2.1. Le componenti principali costruite da S e da R non sono uguali, in generale però sarà chiaro dal contesto quale matrice viene utilizzata.

2.3 Componenti principali per un campione con dati standardizzati

Quando i dati del campione presentano unità di misura o ordini di grandezza differenti, procedendo con le componenti principali ottenute dalla matrice di covarianza l'analisi può diventare fuorviante.

La covarianza, infatti, dipende direttamente dalle unità di misura: variabili con varianza elevata (perché espresse in unità grandi o perché numericamente molto più disperse) contribuiscono in modo sproporzionato al valore del quoziente di Rayleigh. Poiché la PCA seleziona le direzioni che massimizzano la varianza, le variabili con valori numerici assoluti maggiori tendono automaticamente a dominare la costruzione delle componenti principali, anche quando non rappresentano la relazione statistica più rilevante nel dataset.
Per evitare l'insorgere di queste problematiche, si applica la standardizzazione, che rende confrontabili tutte le variabili e consente alle componenti principali di descrivere in modo più fedele le relazioni presenti nei dati.
Al posto della matrice di dati $X \in \mathbb{R}^{n\times p}$ si utilizzerà $Z \in \mathbb{R}^{n\times p}$ tale che:

$$z_i = (x_i - \bar{x_i}^T)D^{-\frac{1}{2}} = \begin{bmatrix} \frac{x_{1i} - \bar{x}_i}{\sqrt{s_{11}}} \\ \frac{x_{2i} - \bar{x}_i}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{ni} - \bar{x}_i}{\sqrt{s_{pp}}} \end{bmatrix}, \quad i = 1, \dots, p,$$

con $D = \text{diag}(s_{11}, \dots, s_{pp})$.

Osservazione 2.3.1. Ogni variabile standardizzata z_i soddisfa:

- *(Varianza campionaria)* $var(z_i) = 1$;
- *(Media campionaria)* $\bar{z}_i = 0$;
- *(Covarianza campionaria)* $cov(z_k, z_i) = \dfrac{s_{ki}}{\sqrt{s_{kk}s_{ii}}} = r_{ki}$.

Definendo la matrice media $\bar{X} \in \mathbb{R}^{n\times p}$ tale che ogni riga è della forma $\bar{x}_1, \dots, \bar{x}_p$;

$$Z = (X - \bar{X})D^{-\frac{1}{2}} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{s_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{s_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{s_{pp}}} \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{x_{11} - \bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{12} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{1p} - \bar{x}_p}{\sqrt{s_{pp}}} \\ \frac{x_{21} - \bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{22} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{2p} - \bar{x}_p}{\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1} - \bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{n2} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{np} - \bar{x}_p}{\sqrt{s_{pp}}} \end{pmatrix}.$$

A questo punto la matrice di covarianza sarà uguale alla matrice di correlazione $S := \frac{1}{n-1} Z^T Z$, e quindi anche le sue autocoppie.
Le componenti principali, in questo caso, saranno combinazioni lineari delle p colonne di Z.

Proposizione 2.3.1. *Data $R \in \mathbb{R}^{p \times p}$ matrice di correlazione e $(\lambda_1, v_1), \ldots, (\lambda_p, v_p)$ autocoppie di R tali che $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$; definendo le componenti principali $y_i = Xa_i = Xv_i$ come sopra, allora:*

1. *(Varianza campionaria) $var(y_i) = \lambda_i \quad \forall i = 1, \ldots, p$;*
2. *(Covarianza campionaria) $cov(y_k, y_i) = 0 \quad \forall i \neq k$;*
3. *(Varianza totale campionaria) $\sum_{i=1}^{p} \lambda_i$.*

2.4 Scelta delle componenti principali

Dal momento in cui la matrice delle componenti principali è data da:

$$Y = Z^* V \quad \text{con} \quad \begin{array}{l} Z \in \mathbb{R}^{n \times p} \text{ matrice dei dati standardizzati,} \\ V \in \mathbb{R}^{p \times p} \text{ matrice degli autovettori della matrice di correlazione;} \end{array}$$

ogni elemento v_{ij} della matrice V dice quanto la variabile originaria standardizzata z_j contribuisca alla componente principale y_i.

Come già osservato precedentemente, ciascuna componente principale può essere interpretata come una combinazione lineare delle variabili iniziali orientata lungo una direzione che massimizza la variabilità spiegata.
Essendo ogni componente principale associata a una direzione che massimizza la variabilità residua del dataset, le componenti possono essere naturalmente ordinate in base alla quantità di varianza che spiegano. È proprio da questa struttura gerarchica che discende l'obiettivo principale della PCA, ovvero quello di sintetizzare p variabili in un numero k di variabili, con $k << p$ in modo da avere una minima perdita di informazione con una grossa riduzione di dati.
Non vi è una procedura standard poichè ci sono molteplici fattori da tenere in considerazione per fare questa scelta. I metodi più comuni e che verranno illustrati e utilizzati in questa tesi sono tre:

1. *(Valutazione grafica)* Si traccia il grafico degli autovalori di S o R, precedentemente messi in ordine decrescente, e si cerca il "gomito", ovvero il cambio di pendenza oltre il quale l'incremento di varianza diventa marginale.

2. *(Percentuale di varianza spiegata)* Si scelgono le prime k componenti principali che spiegano almeno il 60-80% della varianza totale del dataset, ovvero si calcola il vettore rapporto di variabilità e si sceglie il numero di componenti di rv maggiori di una certa soglia 0.6-0.8;

3. *(Autovalori maggiori della media)* Si scelgono i k autovettori (o componenti principali) corrispondenti ai k autovalori maggiori della media degli autovalori stessi (se i dati sono standardizzati la media degli autovalori è 1).

2.5 Interpretazione grafica

Per comprendere appieno le trasformazioni effettuate dall'algoritmo PCA e per interpretarne correttamente i risultati, un ruolo centrale è svolto dalla rappresentazione grafica dei dati. L'osservazione dei grafici di dispersione, infatti,

consente di cogliere aspetti strutturali che non emergerebbero dal solo esame matriciale o numerico, poichè permette di localizzare casi anomali o al contrario identificare zone con maggior concentrazione di dati.
Utile è analizzare tre principali tipi di grafici:

1. *(Tra due variabili originali)* I grafici di dispersione costruiti tra le variabili originali, consentono una prima valutazione della struttura interna del dataset prima di qualsiasi trasformazione, permettendo di individuare correlazioni lineari tra coppie di variabili, ridondanze tra misure, pattern non lineari, outlier (o osservazioni anomale) che si presentano come punti isolati e cluster naturali già presenti nello spazio.

2. *(Tra una variabile originale e una componente principale)* I grafici che mettono in relazione le variabili originali con le componenti principali, consentono di interpretare meglio le nuove coordinate ottenute tramite PCA. Da questo tipo di grafico si può intuire quanto ogni variabile contribuisce a una particolare componente principale, in altre parole, quanto una variabile originale sia ben rappresentata in una specifica componente. In generale, se le nubi di punti appaiono inclinate o orientate, si ha una forte rappresentazione della relazione fra le variabili; al contrario, nubi prevalentemente orizzontali o verticali indicano una rappresentazione debole. Questo passaggio è essenziale per assegnare un significato concreto alle componenti principali.

3. *(Tra due componenti principali)* Gli scatter plot costruiti nello spazio delle componenti principali (ad esempio grafici 2D tra PC1–PC2 o grafici 3D tra PC1–PC2-PC3) sono gli strumenti grafici più utilizzati nell'interpretazione pratica della PCA, infatti, le componenti principali sono ortogonali e non correlate, oltre a conservare la quota massima possibile di varianza del dataset. Questi grafici consentono di visualizzare cluster, individuare outlier e confrontare differenze tra sottogruppi di dati. La visualizzazione nello spazio PC1–PC2, oppure in quello PC1-PC2-PC3, costituisce quindi

un efficace riassunto della struttura dei dati e una base solida per analisi successive, come le tecniche di clustering.

Capitolo 3

Clustering

Il clustering è un processo di raggruppamento di elementi simili, rispetto a determinate caratteristiche, in un insieme di dati. L'obiettivo è individuare una struttura nei dati tale per cui gli oggetti appartenenti allo stesso cluster risultino simili tra loro, ma dissimili da oggetti appartenenti a cluster differenti. Poichè i risultati dipendono sia dall'obiettivo dell'indagine che dal contesto applicativo, è necessario scegliere con attenzione la procedura di raggruppamento e la distanza più adatta per il tipo di dati presi in esame. Questo processo può essere applicato in molti ambiti per aiutare a identificare pattern di raggruppamento e a suddividere le osservazioni in sottogruppi con caratteristiche simili.

3.1 Misure di dissimilarità e distanza

Un primo metodo per formare o separare gruppi di oggetti è quello dei criteri di somiglianza o dissimilarità.

Definizione 3.1.1. Date p variabili binarie, si definisce la tabella di contingenza:

	1	0	totali
1	a	b	$a+b$
0	c	d	$c+d$
totali	$a+c$	$b+d$	$p=a+b+c+d$

a frequenza $1-1$; d frequenza $0-0$; b frequenza $1-0$; c frequenza $0-1$.

Definizione 3.1.2. Sia X un insieme qualunque, una funzione $s : X \times X \to \mathbb{R}$ si dice coefficiente di similarità su X se, per ogni $P, Q \in X$ valgono almeno le seguenti proprietà:

1. *(Simmetria)* $s(P, Q) = s(Q, P)$;
2. *(Non negatività)* $d(P, Q) \geq 0$;
3. *(Massima similarità sull'identità)* $d(P, P) = \max s$;
4. *(Monotonicità crescente)* $s(P, Q) \nearrow$;
5. *(Normalizzazione)* $0 \leq d(P, Q) \leq 1$ (non sempre da soddisfare).

Definizione 3.1.3. Alcuni esempi di coefficienti di similarità sono:

- $s(P, Q) = \dfrac{a}{p}$ stabilisce che P, Q sono simili quando entrambi sono 1;
- *(Simple matching)* $s(P, Q) = \dfrac{a + d}{p}$ stabilisce che P, Q sono simili quando hanno entrambi lo stesso peso;
- *(Coefficiente Jaccard)* $s(P, Q) = \dfrac{a}{a + b + c}$ attribuisce peso nullo al termine $0 - 0$;
- *(Coefficiente Sorensen-Dice)* $s(P, Q) = \dfrac{2a}{2a + b + c}$ attribuisce peso doppio al termine $1 - 1$;
- *(Coefficiente Sokal-Sneath)* $s(P, Q) = \dfrac{2(a + d)}{2(a + d) + b + c}$ attribuisce peso doppio ai termini $1 - 1$ e $0 - 0$;
- *(Coefficiente Rogers-Tanimoto)* $s(P, Q) = \dfrac{a + d}{2(b + c) + a + d}$ attribuisce peso doppio ai termini $1 - 0$ e $0 - 1$;

Definizione 3.1.4. Sia X un insieme qualunque, una funzione $d : X \times X \to \mathbb{R}$ si dice distanza su X se, per ogni $P, Q, R \in X$ valgono le seguenti proprietà:

1. *(Simmetria)* $d(P, Q) = d(Q, P)$;

2. *(Non negatività)* $d(P,Q) \geq 0$;
3. *(Identità degli indiscernibili)* $d(P,Q) = 0 \iff P = Q$;
4. *(Disuguaglianza triangolare)* $d(P,Q) \leq d(P,R) + d(R,Q)$.

Definizione 3.1.5. Siano $x, y \in \mathbb{R}^p$ è possibile definire le seguenti misure di distanza:

- *(Distanza euclidea)* $d(x,y) = \sqrt{(x-y)^T(x-y)}$;
- *(Distanza cityblock)* $d(x,y) = \sum_{i=1}^{p} |x_i - y_i|$;
- *(Distanza cosine)* $d(x,y) = 1 - \dfrac{x^T y}{\|x\| \, \|y\|}$;
- *(Distanza Mahalanobis o statistica)* $d(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$, S matrice di covarianza;
- *(Distanza di Minkowsky)* $d(x,y) = (\sum_{i=1}^{p} |x_i - y_i|^m)^{\frac{1}{m}}$.

Osservazione 3.1.1. Esistono due tipologie di distanze:

- *"within" the group*, che indica quanto sono vicine le osservazioni all'interno di un gruppo, consente di valutarne la coesione e di verificare se una diversa scelta di raggruppamento potrebbe risultare più appropriata;
- *"between" the groups*, che indica quanto sono distanti tra loro i gruppi.

3.2 Metodi gerarchici

I metodi gerarchici di clustering sono una categoria di metodi di analisi multivariata che costruiscono un dendrogramma, ovvero un diagramma ad albero, per rappresentare come le osservazioni o le variabili si uniscono progressivamente in gruppi, senza richiedere a priori il numero di cluster.
Questo gruppo di metodi è ulteriormente suddivisibile in base all'approccio che viene adottato:

- *(Metodi agglomerativi)* Si parte da n cluster singoli, uno per ciascuna osservazione (è possibile farlo anche con le variabili) e a ogni passo si uniscono i due gruppi più simili secondo un criterio di distanza fissato all'inizio. Il processo continua finché tutte le osservazioni formano un unico cluster.

- *(Metodi divisivi)* Si parte da un unico cluster contenente tutte le osservazioni (è possibile farlo anche con le variabili) e lo si suddivide iterativamente in gruppi più piccoli.

Algoritmo 3.2.1. *La tipica procedura in un metodo gerarchico agglomerativo è la seguente:*

1. *Si inizia con n gruppi e una matrice simmetrica* $D \in \mathbb{R}^{n\times n}$ *delle distanze;*

2. *Si determina la coppia di elementi* u, v *più vicini (in termini della distanza scelta) e si forma poi il gruppo* (u, v)*;*

3. *Si aggiorna D, la quale diventerà* $(n-1)\times(n-1)$*, sostituendo alle due righe di u e v una sola riga con le distanze del gruppetto* (u, v) *dagli alti oggetti;*

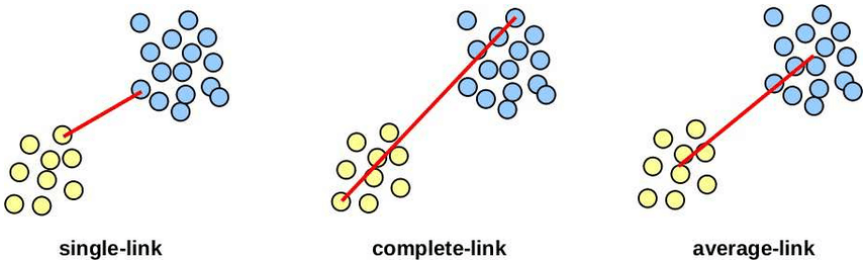
4. *Si ripetono tutti i passaggi a partire dal punto 2, fino a quando* $D \in \mathbb{R}^{1\times 1}$*.*

Tra i metodi gerarchici agglomerativi si possono trovare quelli di connessione o linkage che, in base al tipo di distanza, vengono suddivisi in:

- *(Single Linkage)* La distanza tra due cluster è la minima distanza tra i rispettivi elementi;

- *(Complete Linkage)* La distanza tra due cluster è la massima distanza tra i rispettivi elementi;

- *(Average Linkage)* La distanza tra due cluster è la media delle distanze tra i rispettivi elementi.



Osservazione 3.2.1. Alcune proprietà dei metodi agglomerativi sono:

- Il livello a cui avviene il raggruppamento è importante perchè evidenzia l'effettiva distanza di un elemento dal cluster in cui viene inserito;
- Se la matrice delle distanze D ha minimi uguali con indici diversi, si raggruppano i cluster separatamente:

$$es: \quad D = \begin{pmatrix} 0 & 9 & 3 & 6 & 11 \\ 9 & 0 & 7 & 2 & 10 \\ 3 & 7 & 0 & 9 & 2 \\ 6 & 2 & 9 & 0 & 8 \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix}, \quad \min d_{i,j} = 2$$

si hanno due possibilità $\begin{matrix} i = 5, j = 3 \Rightarrow (3,5) \text{ gruppo} \\ \wedge \\ i = 4, j = 2 \Rightarrow (4,2) \text{ gruppo;} \end{matrix}$

- Se D ha minimi uguali con indici in comune, si raggruppano solo gli oggetti con la stessa distanza:

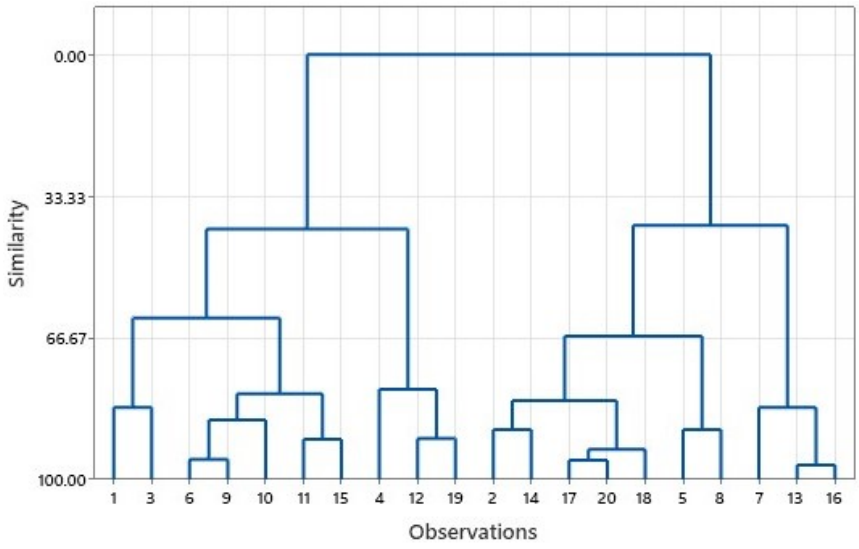
$$es: \quad D = \begin{pmatrix} 0 & 9 & 3 & 6 & 11 \\ 9 & 0 & 2 & 5 & 10 \\ 3 & 2 & 0 & 9 & 2 \\ 6 & 5 & 9 & 0 & 8 \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix}, \quad \min d_{i,j} = 2$$

si hanno due possibilità $\begin{matrix} i = 5, j = 3 \\ i = 3, j = 2 \end{matrix}$ ma $d_{5,2} = 10$

$\Rightarrow (3,5) \dot{\vee} (3,2)$ gruppo, ma non $(2,3,5)$;

- i cluster e i dendrogrammi rimangono inalterati se si usano distanze che mantengono lo stesso ordine.

Definizione 3.2.1. Un dendrogramma è un diagramma ad albero che rappresenta visivamente la disposizione dei cluster prodotti dal clustering gerarchico. È uno strumento cruciale in statistica, analisi dei dati e data science; in particolare quando si ha a che fare con dataset complessi che richiedono l'identificazione di relazioni tra vari punti dati.
Sull'asse delle ascisse del seguente grafico si trovano le osservazioni, mentre sull'asse delle ordinate la distanza. Un aspetto rilevante consiste nell'effettuare un taglio orizzontale, a una certa distanza fissata, in modo da individuare chiaramente i gruppi che emergono dalla struttura gerarchica.



3.3 Metodi non gerarchici

I metodi di clustering non gerarchici, chiamati anche metodi di partizione, sono algoritmi di apprendimento non supervisionato che dividono un insieme di dati in un numero predefinito k di gruppi.
A differenza dei metodi gerarchici non creano una struttura ad albero, ma suddividono direttamente lo spazio in modo che gli elementi appartenenti allo stesso

cluster siano il più possibile simili tra loro, mentre quelli appartenenti a cluster diversi siano il più possibile dissimili.
Un possibile approccio per ottenere questo risultato potrebbe consistere nell'elencare tutti i possibili raggruppamenti in k gruppi costruibili con i dati di partenza, e scegliere come migliore soluzione, quella che ottimizza un determinato criterio predefinito. Sfortunatamente un tale approccio diventerebbe rapidamente inapplicabile, specialmente per grandi dataset, poiché richiederebbe una quantità enorme di tempo macchina e di spazio di memoria. Di conseguenza tutte le tecniche di clustering disponibili sono iterative e operano solo su un numero molto ristretto di enumerazioni.
Di questo gruppo di metodi fa parte l'algoritmo delle K-medie, che opera categorizzando i punti dati in k cluster sulla base di una misura di distanza matematica dal centro di ogni cluster.
L'obiettivo è minimizzare la somma delle distanze tra i punti dati e i cluster assegnati: $\text{SSE} = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$, dove C_j è il cluster j-esimo e μ_j il suo centroide.
Un valore k più alto indica cluster più piccoli con maggiori dettagli, mentre un valore k più basso si traduce in cluster più grandi con meno dettagli.

Definizione 3.3.1. Si dice centroide di un gruppo o cluster la media (come nel caso di K-medie) o la mediana di tutti i punti all'interno del cluster.

Algoritmo 3.3.1. *La tipica procedura per il metodo delle K-medie è la seguente:*

1. *Si suddividono gli oggetti in k gruppi, con k dato in input, e si calcola (in termini della distanza scelta) il centroide di ognuno di essi;*
2. *Si calcola, per ogni osservazione, la distanza dai centroidi di ogni gruppo;*
3. *Si riposiziona ogni oggetto nel cluster con centroide più vicino;*
4. *Si ricalcolano i centroidi dei cluster che hanno acquisito o perso almeno un oggetto;*
5. *Si ripetono tutti i passaggi a partire dal punto 2, fino a quando nessun oggetto cambia più cluster.*

Per rappresentare graficamente i risultati di questo metodo è possibile fare un plot delle prime 2 o 3 componenti principali ottenute con la PCA, evidenziando con diversi colori i gruppi ricavati e con delle $\times$ i centroidi di ogni gruppo.

Osservazione 3.3.1. Per verificare la stabilità del risultato bisognerà applicare più volte l'algoritmo cambiando l'inizializzazione, ovvero il parametro k oppure il tipo di distanza usata.

Osservazione 3.3.2. Il problema della determinazione del valore di k da fornire in input può essere risolto utilizzando il grafico a gomito, che mostra l'andamento dell'errore quadratico totale ("within-cluster sum of squares" - SSE), al variare di k.
Questo valore misura quanto i punti sono vicini al centroide del proprio cluster: valori più bassi indicano una migliore coesione interna.
Aumentando k, SSE tende a diminuire; l'obiettivo è trovare un punto oltre il quale l'incremento di k non porta un miglioramento significativo: questo punto è chiamato "gomito".

Osservazione 3.3.3. Sebbene l'algoritmo delle K-medie sia uno degli algoritmi di clustering più diffusi per la sua semplicità e velocità, non è sempre la scelta migliore a causa della sua pesantezza e lentezza nell'adattarsi ad una grande mole di dati. Infatti ha una complessità computazionale $O(r \cdot n \cdot k \cdot i \cdot p)$ che cresce linearmente con n (numero di osservazioni), k (numero di cluster), i (numero di iterazioni), r (numero di riavvii) e p (numero di variabili), quindi su dataset molto grandi può risultare computazionalmente costoso, soprattutto senza ottimizzazioni.

3.4 Confronto tra metodi

Caratteristica	**Gerarchico Agglomerativo (Complete Linkage)**	**K-medie**
Tipo di algoritmo	Deterministico; costruisce una gerarchia di fusione tra cluster	Stocastico; la soluzione dipende dall'inizializzazione
Input richiesti	Non richiede il numero di cluster a priori; usa solo la matrice delle distanze	Richiede a priori il numero di cluster k
Struttura prodotta	Dendrogramma multilivello	Partizione unica in k gruppi
Forma dei cluster	Cluster di forme arbitrarie	Cluster compatti e sferici
Sensibilità agli outlier	Moderata (complete è più robusto di single)	Alta: gli outlier influenzano i centroidi
Costo computazionale	Elevato: $O(n^2)$	Più efficiente: circa $O(n \cdot k \cdot p)$ per iterazione, senza considerare i riavii
Interpretazione	Molto intuitiva via dendrogramma	Meno intuitiva ad alte dimensioni
Stabilità della soluzione	Alta stabilità	Bassa stabilità (richiede più riavvii)

Capitolo 4

Applicazione a un dataset

4.1 Presentazione del dataset

I dati analizzati in questa tesi sono stati forniti da Vittorio Coloretti, dottorando presso il Dipartimento di Scienze per la Qualità della Vita (Unibo). Tale dataset è costituito da un campione di 73 individui, per ciascuno dei quali sono state raccolte 47 tra variabili antropometriche e prestative relative ai quattro diversi stili del nuoto.

- Sesso *(M=0/F=1)*

dati antropometrici

- Età *(anni)*
- Altezza *(cm)*
- Peso *(kg)*
- Altezza in stream *(cm)*
- Distanza biacromiale *(cm)*
- Larghezza delle spalle in stream *(cm)*
- BMI *(kg/cm^2)*

dati prestativi

- PB sui 50 metri *(s)*
- Punti FINA relativi a quel personale
- Ft *(N)*
- Vmax *(m/s)*
- Fp *(N)*
- Ka *(N)*
- Kp *(N)*
- Kcin
- SR Vmax *(cicli/min)*
- SL Vmax *(m)*
- SR Ft *(cicli/min)*

Definizione 4.1.1. La distanza biacromiale è la distanza in linea retta tra i due acromion, ossia due processi ossei della scapola che formano la parte più alta e posteriore delle spalle, collegandosi alla clavicole.

Definizione 4.1.2. La posizione di "streamline" (o semplicemente "stream"), nel nuoto, è una posizione del corpo idrodinamica e affusolata assunta durante le fasi subacquee come a seguito di partenza o virate, per minimizzare l'attrito e massimizzare la velocità. Si ottiene allineando braccia, tronco e gambe, con le braccia unite sopra la testa e i gomiti stretti intorno alle orecchie: il corpo è il più possibile orizzontale con la testa allineata.

Definizione 4.1.3. Il BMI o Indice di Massa Corporea è un valore che indica il rapporto tra il peso e l'altezza di una persona, calcolato dividendo il peso in chilogrammi per il quadrato dell'altezza in metri. Non è utile a valutare la quantità di grasso, ma fornisce una rapida e generale stima dello stato nutrizionale, classificando il peso come sottopeso, normopeso, sovrappeso o obeso.

Definizione 4.1.4. I punti FINA sono un sistema di valutazione, per confrontare le prestazioni nel nuoto, che assegna un punteggio più alto a chi si avvicina o supera i tempi base stabiliti annualmente da World Aquatics. Il punteggio dipende dal tempo ottenuto rispetto a un tempo di base prestabilito, che è basato sui record mondiali più recenti e si differenzia tra vasca corta e vasca lunga.

Definizione 4.1.5. La forza propulsiva (Fp) è la forza generata dal nuotatore per vincere la resistenza idrodinamica e produrre avanzamento.

Definizione 4.1.6. La forza al full tethered (Ft) è la massima forza propulsiva che il nuotatore è in grado di esprimere durante una nuotata stazionaria, cioè vincolata a un punto fisso tramite un cavo.

Definizione 4.1.7. In generale il drag rappresenta la forza di resistenza che il fluido esercita sul corpo che si muove al suo interno e tende a crescere all'aumentare della forza propulsiva e della velocità.
Nel nuotatore agiscono diversi tipi di drag:

- *(Drag attivo - Ka)* Resistenza idrodinamica che il nuotatore incontra mentre si muove per generare propulsione, influenzata da tecnica, posizione del corpo,turbolenze e movimenti di braccia e gambe.

- *(Drag passivo - Kp)* Resistenza idrodinamica sull'atleta immobile in posizione streamline, determinata principalmente da forma, superficie frontale e profilo del corpo.

- *(Drag cinetico - Kcin)* Resistenza creata dalle turbolenze generate dal movimento degli arti, dipendente da variazioni posturali e dal modo in cui i segmenti attraversano il fluido. Esso è la somma tra Ka e Kp.

Definizione 4.1.8. Il ciclo di bracciata nel nuoto è la sequenza completa di movimenti che una singola mano compie dall'entrata in acqua fino al suo successivo rientro. Si suddivide in quattro fasi principali: appoggio/presa, trazione, spinta e recupero. A stile libero e a dorso un ciclo equivale a due bracciate, mentre a rana e delfino a una bracciata.

Definizione 4.1.9. La frequenza di bracciata (SR) è il numero di cicli di bracciata completati in un minuto (o un altro intervallo di tempo prestabilito).

Definizione 4.1.10. L'ampiezza di bracciata (SL) è la distanza coperta da un ciclo completo di bracciata (dall'ingresso della mano in acqua all'ingresso successivo della stessa mano).

Il dataset considerato, di dimensioni 73×47, presenta diversi valori mancanti. La gestione di questi valori rappresenta una delle problematiche più rilevanti e, al tempo stesso, più delicate nell'ambito della statistica applicata e dell'analisi di dati (si vedano [4] e [3]).
Si parla di dati mancanti quando uno o più valori all'interno di un set di dati non risultano disponibili o, in altre parole, quando, per una determinata variabile o osservazione, l'informazione attesa non è presente. Le cause possono essere molteplici: errori nella fase di raccolta o digitalizzazione dei dati, risposte omesse nei questionari o nei sondaggi, malfunzionamenti di strumenti o sensori

durante la registrazione di misure, perdita accidentale o danneggiamento di archivi informatici, rifiuto di partecipazione da parte dei soggetti o incapacità di completare la procedura sperimentale.
Qualunque sia la motivazione, la presenza di valori mancanti può compromettere in modo significativo la qualità delle analisi statistiche successive, influenzando la bontà delle stime, la validità delle inferenze, la generalizzabilità dei risultati e la capacità predittiva dei modelli.
La letteratura classifica i meccanismi di generazione della mancanza in tre principali categorie:

- *(MCAR – Missing Completely At Random)* La mancanza avviene in modo completamente casuale, senza alcuna relazione con le variabili osservate né con i valori mancanti stessi.

- *(MAR – Missing At Random)* La probabilità che un valore sia mancante dipende dai valori osservati di altre variabili, ma non dai valori mancanti stessi.

- *(MNAR – Missing Not At Random)* La mancanza dipende direttamente dal valore non osservato, cosa che rende MNAR il meccanismo più complesso e problematico.

Oltre al meccanismo, i dati mancanti possono essere caratterizzati da:

- *(Tasso di mancanza)* Proporzione di celle mancanti sul totale del dataset. Si considera in genere mancanza lieve se inferiore al 10%, moderata tra 10% e 25%, elevata tra 25% e 50% e eccessiva oltre il 50%.

- *(Modello di mancanza)* Descrive la distribuzione dei valori mancanti nelle variabili e può essere categorizzato come univariato (mancano dati in una sola variabile), multivariato (mancano dati in più variabili), monotono (mancano dati in una direzione del dataset), non monotono e connesso (i dati completi sono raggiungibili tramite movimenti orizzontali o verticali all'interno di un dataset tabulare).

Le strategie per affrontare il problema dei dati mancanti si articolano in tre grandi classi di metodi: eliminazione, imputazione e apprendimento della rappresentazione.
I metodi di eliminazione rappresentano l'approccio più semplice e immediato e si suddividono ulteriormente in metodi di eliminazione listwise, che rimuovono tutte le osservazioni le cui righe di variabili contengono almeno un valore mancante, e metodi di eliminazione pairwise, che escludono le osservazioni a cui mancano alcuni dati solo durante l'analisi che le comprende. Questi approcci sono validi solo se i dati sono MCAR e il tasso di mancanza è contenuto.
I metodi di imputazione si propongono di sostituire i valori mancanti con stime plausibili, preservando la struttura multivariata del dataset e si dividono in metodi di imputazione singola, multipla e basata su decomposizione.
Nel caso di imputazione singola, le stime possono essere ottenute tramite regole statistiche, ad esempio media, mediana o moda dei valori non mancanti della colonna; tramite modelli di regressione lineare o logistica; tramite sostituzione con valori non mancanti, ad esempio il valore dell'osservazione precedente o successiva (LOCF o NOBC) o il valore del vicino più prossimo in base alla distanza scelta (KNN); tramite modelli probabilistici, ad esempio l'algorito iterativo EM, che stima i valori mancanti massimizzando la funzione di verosimiglianza sotto uno specifico modello probabilistico; oppure tramite modelli predittivi, ad esempio alberi decisionali e foreste casuali. Quando la quantità di dati mancanti cresce, l'imputazione singola può risultare insufficiente. In questi casi vengono utilizzate tecniche di imputazione multipla, nelle quali si generano più versioni complete del dataset, ciascuna con imputazioni leggermente differenti, e si combinano i risultati per ottenere stime robuste. Nel caso di imputazione basata su decomposizione, si utilizzano versioni modificate dell'algoritmo PCA o della SVD per stimare i valori mancanti in base alle componenti principali o alle strutture latenti identificate.
I metodi di apprendimento della rappresentazione sono invece tecniche più sofisticate, introdotte con l'evoluzione dell'intrelligenza artificiale, che si basano su reti neurali profonde, autoencoder o modelli probabilistici latenti, apprendendo

automaticamente strutture e relazioni complesse nei dati grezzi poi utilizzate per migliorare la qualità e l'accuratezza delle imputazioni.
Ai fini dell'analisi presentata in questa tesi, risulta fondamentale comprendere come gestire l'incompletezza del dataset nell'ambito della PCA.
Diversi approcci sono stati proposti nel tempo per risolvere questo problema, come l'imputazione basata su SVD (I-SVD), in cui le voci mancanti vengono riempite e aggiornate a ogni iterazione della SVD fino alla convergenza, e l'adattamento dell'algoritmo Nonlinear Iterative Partial Least Squares (NIPALS), in grado di saltare le voci mancanti durante la stima dei minimi quadrati di scores e loadings. Tuttavia, sono state segnalate alcune limitazioni per entrambi gli approcci: la convergenza dell'algoritmo I-SVD può essere molto lenta per set di dati con un'alta percentuale di missings e, quando si utilizza NIPALS, le proprietà di ortogonalità tra scores e loadings potrebbero essere perse.
Per risolvere questi problemi, è possibile utilizzare un algoritmo denominato Orthogonalized-Alternating Least Squares (O-ALS), ovvero un algoritmo di minimi quadrati alternati che stima scores e loadings, soggetti al vincolo di ortogonalizzazione di Gram-Schmid (si vedano [2] per un'analisi completa sugli approcci da usare nella PCA con valori mancanti e [1] per un confronto tra i tre metodi appena citati).
L'algoritmo O-ALS inizia con una stima, solitamente casuale o basata su una prima approssimazione, della matrice dei loadings e successivamente procede in modo iterativo secondo uno schema alternato (alternating scheme):

- stima riga per riga della matrice di loadings con i minimi quadrati, applicando a ogni colonna della matrice stessa il vincolo di ortogonalizzazione di Gram-Schmidt;

- stima colonna per colonna della matrice di scores con i minimi quadrati, applicando a ogni riga della matrice stessa il vincolo di ortogonalizzazione di Gram-Schmidt;

e prosegue fino a convergenza.

4.2 Analisi con dataset ristretto

L'analisi è stata condotta inizialmente su un sottoinsieme di 42 individui per i quali sono presenti sia le variabili antropometriche sia quelle prestative nello stile libero. Il dataset considerato ha quindi dimensione 42×19.

	sesso	età($anni$)	alt.(cm)	peso(kg)	alt.stream(cm)	dist.biac.(cm)	larg.spal.stream(cm)	BMI(kg/cm^2)
s1	0	22	177	73	243	36	35	23.3
s2	0	19	177	70	240	37	38	22.3
s3	0	22	180	78	248	35	37	24.1
s4	0	20	190	87	263	41	40	24.1
s5	0	22	183	76	250	39	39	22.7
s6	0	21	178	83	246	37	39	26.2
s7	1	18	169	55	228	32	31	19.3
s8	1	21	161	56	219	33	35	21.6
s9	1	13	163	55	224	31	34	20.7
s10	1	15	163	54	225	31	32	20.3
s11	1	18	162	57	222	31	33	21.7
s12	1	16	160	55	225	33	32	21.5
s13	0	14	176	69	240	34	33	22.3
s14	1	13	149	39	205	31	33	17.6
s15	0	27	185	83	256	37	40	24.3
s16	0	16	181	74	243	35	38	22.6
s17	1	13	171	53	229	32	34	18.1
s18	1	13	165	60	232	35	32	22.0
s19	1	14	164	60	222	32	32	22.3
s20	1	18	168	60	234	32	36	21.3
s21	1	18	168	61	227	36	36	21.6
s22	1	23	178	72	243	36	38	22.7
s23	1	23	177	68	238	35	33	21.7
s24	1	21	173	63	234	36	34	21.0
s25	1	23	182	70	249	37	35	21.1
s26	1	18	175	65	237	35	34	21.2
s27	0	18	187	76	258	38	35	21.7
s28	1	18	165	60	222	33	36	22.0
s29	1	15	165	56	229	33	33	20.6
s30	1	20	169	63	227	32	32	22.1
s31	1	20	170	58	233	32	35	20.1
s32	1	24	161	57	220	32	33	22.0
s33	0	19	183	75	244	36	38	22.4
s34	0	23	183	90	247	38	39	26.9
s35	0	20	181	69	251	36	35	21.1
s36	1	20	169	62	223	32	34	21.7
s37	0	24	168	63	229	34	37	22.3
s38	0	22	189	73	261	38	36	20.4
s40	0	21	193	81	265	39	37	21.7
s41	0	26	181	78	253	37	37	23.8
s53	0	27	184	73	256	38	34	21.6
s58	1	25	172	63	231	31	36	21.3

	PB50(s)	FINApoints	Ft(N)	Vmax(m/s)	Fp(N)	Ka(kg/m)	Kp(kg/m)	Kcin	SRvmax($cicli/min$)	SLvmax(m)	SRFt($cicli/min$)
s1	24.76	539	148.9	1.82	142.8	43.6	22.8	1.91	58.0	1.88	59.4
s2	25.08	519	134.8	1.81	134.8	43.0	23.7	1.81	59.5	1.83	54.7
s3	24.65	547	158.8	1.81	147.5	48.9	29.8	1.64	53.7	2.02	53.5
s4	24.10	585	192.0	1.73	178.5	56.6	29.7	1.91	54.0	1.92	56.2
s5	24.32	569	180.7	1.92	153.5	47.9	26.3	1.82	57.1	2.02	60.0
s6	25.34	503	169.2	1.71	152.7	58.0	27.1	2.14	52.5	1.95	51.9
s7	28.00	549	105.5	1.58	95.3	40.4	21.3	1.90	58.6	1.62	62.8
s8	28.50	520	94.9	1.45	88.1	45.0	21.7	2.07	57.3	1.52	55.8
s9	29.65	462	96.1	1.58	94.9	40.9	27.4	1.49	54.4	1.74	48.7
s10	32.49	337	101.0	1.59	91.1	39.2	20.0	1.96	56.0	1.70	57.0
s11	26.52	646	113.0	1.68	107.5	39.2	20.6	1.90	57.4	1.76	67.0
s12	26.13	675	116.1	1.64	113.8	41.7	20.2	2.06	57.0	1.73	57.6
s13	23.83	600	159.5	1.85	157.1	45.5	32.1	1.42	65.8	1.69	61.9
s14	30.75	414	86.6	1.54	83.4	38.7	17.6	2.20	53.1	1.74	55.3
s15	23.20	656	191.8	1.82	180.1	55.0	24.5	2.24	61.9	1.76	60.7
s16	23.59	624	166.4	1.84	147.5	45.8	28.2	1.62	56.2	1.96	55.4
s17	26.30	662	128.7	1.70	115.1	42.6	20.8	2.05	57.0	1.79	62.2
s18	28.71	509	106.9	1.59	99.6	42.9	23.1	1.86	56.8	1.68	51.5
s19	28.06	545	103.8	1.57	99.9	43.4	22.4	1.94	54.9	1.72	54.1
s20	25.81	701	108.6	1.58	102.4	42.4	23.1	1.84	52.0	1.82	54.6
s21	27.41	585	126.2	1.61	122.1	49.8	21.5	2.32	54.1	1.79	53.6
s22	24.77	793	122.9	1.67	108.3	45.6	24.9	1.83	51.6	1.94	49.8
s23	26.99	613	89.2	1.63	89.0	35.7	24.5	1.46	58.5	1.67	55.0
s24	27.70	567	96.7	1.70	91.4	34.6	22.3	1.55	58.7	1.74	55.3
s25	28.00	549	88.9	1.66	92.0	35.1	24.2	1.45	53.2	1.87	49.1
s26	27.96	551	107.1	1.64	103.0	42.0	17.4	2.41	56.2	1.75	52.6
s27	24.83	535	155.3	1.81	144.0	46.8	23.5	1.99	53.0	2.05	53.1
s28	27.28	593	89.8	1.55	95.5	36.8	19.8	1.86	52.2	1.78	52.5
s29	26.46	650	107.5	1.68	93.5	36.7	17.5	2.10	55.6	1.81	52.0
s30	26.41	654	120.8	1.61	102.2	44.4	23.1	1.92	50.2	1.92	50.5
s31	29.52	468	91.4	1.53	89.3	41.6	22.9	1.82	46.5	1.97	45.6
s32	26.43	653	109.1	1.64	109.0	42.2	22.3	1.89	58.6	1.68	59.1
s33	22.92	680	180.1	1.91	160.8	46.3	22.9	2.02	62.2	1.84	60.1
s34	23.50	631	174.3	1.87	175.5	50.9	25.3	2.01	59.1	1.90	60.3
s35	26.00	466	125.7	1.71	103.2	38.3	25.0	1.53	55.2	1.86	53.6
s36	28.27	533	91.7	1.53	92.8	40.9	23.2	1.76	52.9	1.74	48.0
s37	26.97	417	103.2	1.61	102.9	42.5	22.3	1.91	57.0	1.69	60.0
s38	23.20	656	164.5	1.83	147.7	46.3	23.8	1.95	52.9	2.08	52.8
s40	23.80	608	163.1	1.85	148.0	45.6	26.6	1.71	54.9	2.02	59.0
s41	24.10	585	160.2	1.86	160.4	46.1	28.0	1.65	62.0	1.80	61.6
s53	24.50	557	134.6	1.80	131.7	40.6	24.2	1.68	52.6	2.05	52.5
s58	27.20	599	105.3	1.58	93.0	40.1	23.0	1.74	52.8	1.80	50.9

Dapprima è stata calcolata la matrice di correlazione per valutare le relazioni lineari tra le variabili e successivamente applicata la Principal Component Analysis (PCA) allo scopo di ridurre la dimensionalità del dataset e individuare le

componenti principali maggiormente rappresentative.
In seguito, i risultati ottenuti sono stati impiegati per eseguire un'analisi di clustering mediante algoritmi di tipo Complete Linkage e K-medie, con l'obiettivo di individuare possibili raggruppamenti omogenei tra i soggetti.
Per evitare che l'unica variabile binaria del dataset (il sesso) influenzasse l'analisi, l'intera procedura è stata ripetuta separando i dati per sesso. Sono stati quindi analizzati distintamente i sottoinsiemi uomini e donne, applicando in ciascun caso le fasi di calcolo delle correlazioni, PCA e clustering.
Infine il dataset è stato anche suddiviso per tipologia di variabili: dati antropometrici e dati prestativi (stile libero). Le strutture di cluster ottenute nei due casi sono state confrontate per evidenziare differenze o somiglianze tra la configurazione antropometrica e quella prestativa dei soggetti.

4.2.1 Analisi della matrice di correlazione

Dopo aver standardizzato i dati è stata calcolata la matrice di correlazione associata. In rosso sono evidenziati i valori di alta correlazione, in blu di bassa correlazione.

	sesso	età	altezza	peso	altezza stream	dist.biac.	larg.spal. stream	BMI	PB 50	FINApoints	Ft	Vmax	Fp	Ka	Kp	Kcin	SR vmax	SL vmax	SR Ft
sesso	1.0000	-0.3792	-0.7484	-0.7887	-0.7675	-0.7268	-0.6535	-0.5593	0.7612	0.0303	-0.8422	-0.8336	-0.8472	-0.5827	-0.6200	0.1282	-0.3148	-0.5555	-0.3072
età	-0.3792	1.0000	0.5109	0.5642	0.4966	0.4597	0.4864	0.4689	-0.4339	0.2005	0.2792	0.3003	0.3132	0.2141	0.2548	-0.1435	0.0165	0.2924	0.0362
altezza	-0.7484	0.5109	1.0000	0.8996	0.9692	0.8557	0.6406	0.4684	-0.7564	0.2580	0.7503	0.7701	0.7303	0.4690	0.6052	-0.2406	0.1019	0.6884	0.0548
peso	-0.7887	0.5642	0.8996	1.0000	0.8829	0.8466	0.7562	0.8019	-0.7874	0.2559	0.8216	0.7546	0.8380	0.6647	0.6775	-0.1432	0.1766	0.5991	0.1077
altezza stream	-0.7675	0.4966	0.9692	0.8829	1.0000	0.8768	0.6115	0.4701	-0.7403	0.2153	0.7577	0.7663	0.7397	0.4859	0.6083	-0.2255	0.0827	0.7030	0.0540
dist.biac.	-0.7268	0.4597	0.8557	0.8466	0.8768	1.0000	0.6617	0.5312	-0.6779	0.1550	0.7187	0.7152	0.7273	0.5260	0.4845	-0.0558	0.1465	0.5879	0.0743
larg.spal.stream	-0.6535	0.4864	0.6406	0.7562	0.6115	0.6617	1.0000	0.6311	-0.6284	0.2008	0.6974	0.5413	0.6998	0.6765	0.4865	0.0706	0.0421	0.4930	0.0616
BMI	-0.5593	0.4689	0.4684	0.8019	0.4701	0.5312	0.6311	1.0000	-0.5752	0.1979	0.6148	0.4673	0.6730	0.6768	0.5551	-0.0029	0.2195	0.2653	0.1186
PB 50	0.7612	-0.4339	-0.7564	-0.7874	-0.7403	-0.6779	-0.6284	-0.5752	1.0000	-0.6178	-0.8439	-0.8385	-0.8405	-0.5591	-0.5488	0.0805	-0.3276	-0.5528	-0.3470
FINApoints	0.0303	0.2005	0.2580	0.2559	0.2153	0.1550	0.2008	0.1979	-0.6178	1.0000	0.3047	0.3005	0.2863	0.1729	0.0869	0.0558	0.1252	0.1943	0.1799
Ft	-0.8422	0.2792	0.7503	0.8216	0.7577	0.7187	0.6974	0.6148	-0.8439	0.3047	1.0000	0.8620	0.9747	0.8043	0.6061	0.0964	0.3339	0.5659	0.3858
Vmax	-0.8336	0.3003	0.7701	0.7546	0.7663	0.7152	0.5413	0.4673	-0.8385	0.3005	0.8620	1.0000	0.8571	0.4407	0.5312	-0.1533	0.4568	0.5936	0.4211
Fp	-0.8472	0.3132	0.7303	0.8380	0.7397	0.7273	0.6998	0.6730	-0.8405	0.2863	0.9747	0.8571	1.0000	0.7984	0.6219	0.0781	0.3915	0.5099	0.4083
Ka	-0.5827	0.2141	0.4690	0.6647	0.4859	0.5260	0.6765	0.6768	-0.5591	0.1729	0.8043	0.4407	0.7984	1.0000	0.5196	0.3684	0.0923	0.3625	0.1704
Kp	-0.6200	0.2548	0.6052	0.6775	0.6083	0.4845	0.4865	0.5551	-0.5488	0.0869	0.6061	0.5312	0.6219	0.5196	1.0000	-0.5907	0.1819	0.3826	0.0463
Kcin	0.1282	-0.1435	-0.2406	-0.1432	-0.2255	-0.0558	0.0706	-0.0029	0.0805	0.0558	0.0964	-0.1533	0.0781	0.3684	-0.5907	1.0000	-0.0676	-0.1031	0.1233
SR vmax	-0.3148	0.0165	0.1019	0.1766	0.0827	0.1465	0.0421	0.2195	-0.3276	0.1252	0.3339	0.4568	0.3915	0.0923	0.1819	-0.0676	1.0000	-0.4432	0.7433
SL vmax	-0.5555	0.2924	0.6884	0.5991	0.7030	0.5879	0.4930	0.2653	-0.5528	0.1943	0.5659	0.5936	0.5099	0.3625	0.3826	-0.1031	-0.4432	1.0000	-0.2497
SR Ft	-0.3072	0.0362	0.0548	0.1077	0.0540	0.0743	0.0616	0.1186	-0.3470	0.1799	0.3858	0.4211	0.4083	0.1704	0.0463	0.1233	0.7433	-0.2497	1.0000

Osservando la matrice di correlazione, emergono alcune relazioni lineari marcate tra le variabili.

Le variabili antropometriche mostrano in generale correlazione positiva ed elevata, a conferma del fatto che le grandezze corporee tendono a crescere in maniera proporzionale. Ad esempio la correlazione tra 'altezza' e 'peso' è pari a *0.8996* e addirittura quella tra 'altezza' e 'altezza in stream' raggiunge un valore di *0.9692*.

Un altro aspetto interessante da osservare riguarda la variabile 'sesso' che, in accordo con quanto ci si aspetterebbe, risulta essere negativamente correlata con tutte le variabili antropometriche. Infatti, assumendo la codifica binaria 0=uomo e 1=donna, una correlazione negativa implica che nel passaggio da 0 a 1, i valori di 'altezza', 'peso', 'altezza in stream', 'distanza biacromiale', 'larghezza spalle in stream' e 'BMI' tendano a diminuire. In altre parole, nel campione analizzato, le donne presentano mediamente valori inferiori rispetto agli uomini per queste caratteristiche fisiche. Anche rispetto alle variabili prestative il 'sesso' mostra correlazioni negative, confermando che gli uomini ottengono mediamente prestazioni migliori, ovvero nuotano più velocemente. Fanno eccezione la variabile 'PB sui 50 metri', con cui la correlazione è positiva (*0.7612*) poiché gli uomini nuotano tempi più bassi e le variabili 'FINA points' e 'Kcin', le cui correlazioni con il 'sesso' risultano essere deboli (rispettivamente *0.0303* e *0.1282*), non fornendo quindi informazioni significative.

Osservando ora le variabili prestative si possono notare alcuni gruppi con correlazioni significative, coerenti con l'andamento generale delle performance.

In particolare 'PB sui 50 metri' e 'Vmax' mostrano una forte correlazione negativa (*-0.8385*), in linea con il fatto che tempi minori corrispondono a velocità maggiori; il gruppo di variabili composto da 'Fp', 'Ft', 'Vmax' risulta essere ben correlato positivamente, indicando che queste tre variabili tendono ad aumentare insieme.

Di particolare interesse è anche l'analisi delle tre variabili di 'drag' (quello attivo (Ka), quello passivo, (Kp), e quello cinetico (Kcin)), che si notano (in particolare Ka e Kp) avere una forte correlazione positiva con la 'forza al tethered' (Ft) e con la 'forza propulsiva' (Fp), coerentemente con l'aspetto biomeccanico del nuoto. Inoltre è possibile notare che nel campione considerato, i soggetti

più alti (ka-altezza: *0.4690*, kp-altezza: *0.6052*) e pesanti (ka-peso: *0.6647*, kp-peso: *0.6775*) hanno un maggior drag, ovvero sono soggetti a una maggiore resistenza idrodinamica. Questo aspetto è ulteriormente evidenziato dalle correlazioni negative elevate tra 'Ka' e 'Kp' con il 'sesso' (ka-sesso: *-0.5827*, kp-sesso: *-0.6200*), che indicano come i soggetti maschili tendano ad avere un 'drag' generalmente maggiore rispetto a quelli femminili.
Altra correlazione incrociata interessante tra variabili antropometriche e prestative è quella tra soggetti alti e pesanti e soggetti con 'forza propulsiva' e 'velocità massima' maggiori (altezza-Fp: *0.7303*, altezza-Vmax: *0.7701* e peso-Fp: *0.8380*, peso-Vmax: *0.7546*), evidenziando come la maggiore massa corporea si associ spesso a una maggiore potenza nonostante la resistenza idrodinamica sia maggiore.
Infine si evidenzia una correlazione negativa moderata (*-0.4432*) tra le variabili 'SR vmax' e 'SL vmax', a conferma del fatto che un aumento della frequenza di bracciata comporta una riduzione della lunghezza della stessa.

4.2.2 Grafici di dispersione tra le variabili iniziali

Si procede ora all'esame dettagliato di alcuni grafici di dispersione tra le variabili, analizzandone le principali caratteristiche.

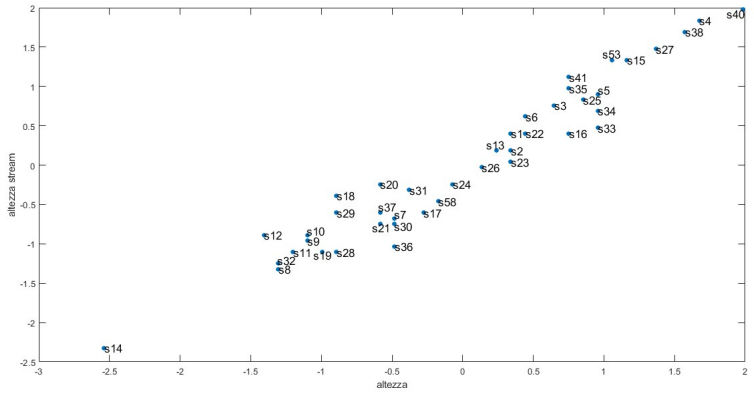


Figura 4.1: Relazione tra 'altezza' e 'altezza in stream'

In Figura 4.1 si ossserva che la correlazione tra le due variabili (0.9692) è diretta o positiva, cioè all'aumentare di una anche l'altra aumenta, e forte. Si nota infatti che l'andamento dei punti rappresentati è approssimabile a una retta crescente. Individui più alti tendono ad avere anche una maggiore estensione del corpo nella posizione di scivolamento ("streamline"). Emerge che l'osservazione s14 è distante dal resto dei punti. Ciò sta a indicare un comportamento anomalo, confermato dai dati presi in esame: il soggetto è una donna con età molto minore rispetto al resto del campione e con misurazioni conseguentemente basse.

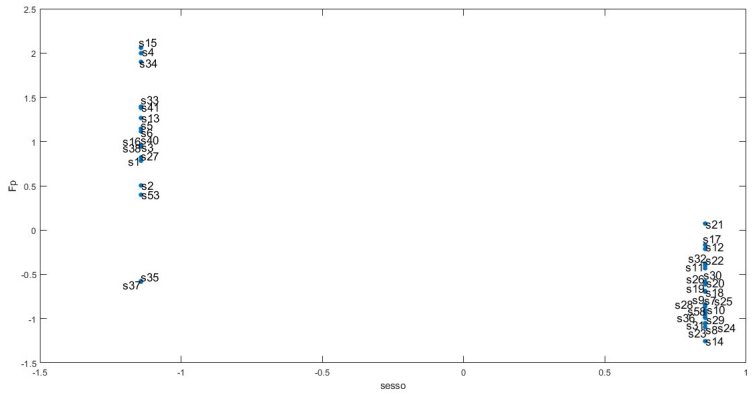


Figura 4.2: Relazione tra 'sesso' e 'forza propulsiva'

In Figura 4.2 si osserva che la correlazione tra queste due variabili (-0.8472) è inversa o negativa, cioè all'aumentare della variabile sesso diminuisce la forza propulsiva, e forte. Essendo la variabile sesso binaria (0=uomini, 1=donne), passando da 0 a 1, ovvero da uomo a donna, la forza propulsiva tende a diminuire. Nel grafico si osservano infatti due gruppi distinti di punti: quelli con ascissa minore (uomini) presentano valori di forza propulsiva mediamente più elevati (ordinata maggiore), mentre quelli con ascissa maggiore (donne) si collocano su valori inferiori (ordinata minore). Questo andamento conferma la differenza di forza muscolare tra i due sessi, già nota in letteratura per quanto riguarda la produzione di forza in acqua.

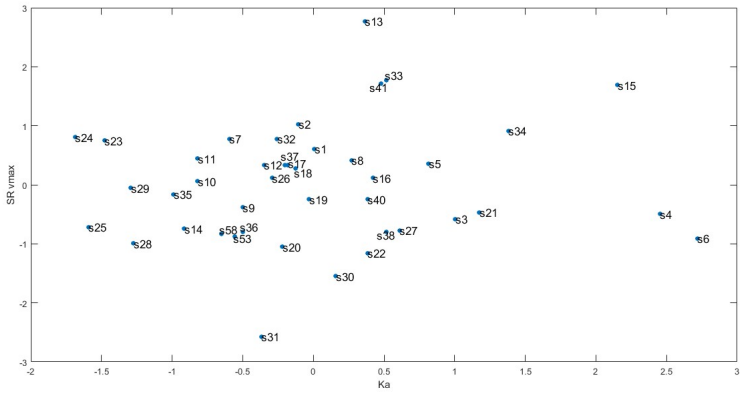


Figura 4.3: Relazione tra 'drag attivo' e 'frequenza di bracciata'

In Figura 4.3 si osserva che la correlazione tra queste due variabili (0.0923) è diretta o positiva e debole. Ciò significa che, pur esistendo una leggera tendenza all'aumento della frequenza di bracciata con l'aumentare del drag attivo, la relazione è poco significativa. Il grafico mostra infatti una nuvola di punti diffusa e priva di una direzione predominante, indice del fatto che la frequenza di bracciata non dipende in modo lineare dal drag attivo, ma è probabilmente influenzata da altri fattori biomeccanici o tecnici (ad esempio la coordinazione o la potenza specifica degli arti superiori).

4.2.3 Analisi delle componenti principali

La presenza di numerosi valori elevati, evidenziati in rosso nella matrice di correlazione Figura 4.2.1, è segnale che molte variabili sono fortemente correlate tra loro. Si parla in questo caso di ridondanza informativa. Questa espressione sta ad indicare che due o più variabili portano informazioni molto simili tra loro, ovvero la loro variazione viene in gran parte spiegata da un'unica "direzione".
Nel contesto dell'analisi di dati, il fatto di avere ridondanza informativa rende l'analisi più complessa, sprecando capacità computazionale e ostacolando l'interpretazione; ciò giustifica l'uso della PCA.
Dopo aver standardizzato il dataset, vengono calcolati la matrice di correlazione e i suoi autovettori e autovalori, per poi dedurre le componenti principali.

```
[n,p]=size(X)     % n=42 (numero righe), p=19 (numero colonne)
Z = zscore(X);    % Z è la matrice 42x19 dei dati standardizzati
R = corr(Z);      % R è la matrice 19×19 di correlazione di Z
[V,D] = eig(R);   % V è la matrice 19×19 con gli autovettori di R
                  % D è la matrice diagonale 19×19 che sulla
                  % diagonale ha i relativi autovalori di R
lambda=diag(D);   % metto gli autovalori in un vettore colonna
[lambda_sorted,idx]=sort(lambda,'descend');
                  % ordino gli autovalori in ordine decrescente
                  % e metto in memoria gli indici
V=V(:,idx);       % ordino le colonne di V in modo coerente
Y=Z*V             % matrice delle componenti principali
```

Attraverso i tre metodi descritti in Sezione 2.4 si deduce il numero necessario delle componenti principali.

1. (Valutazione grafica)

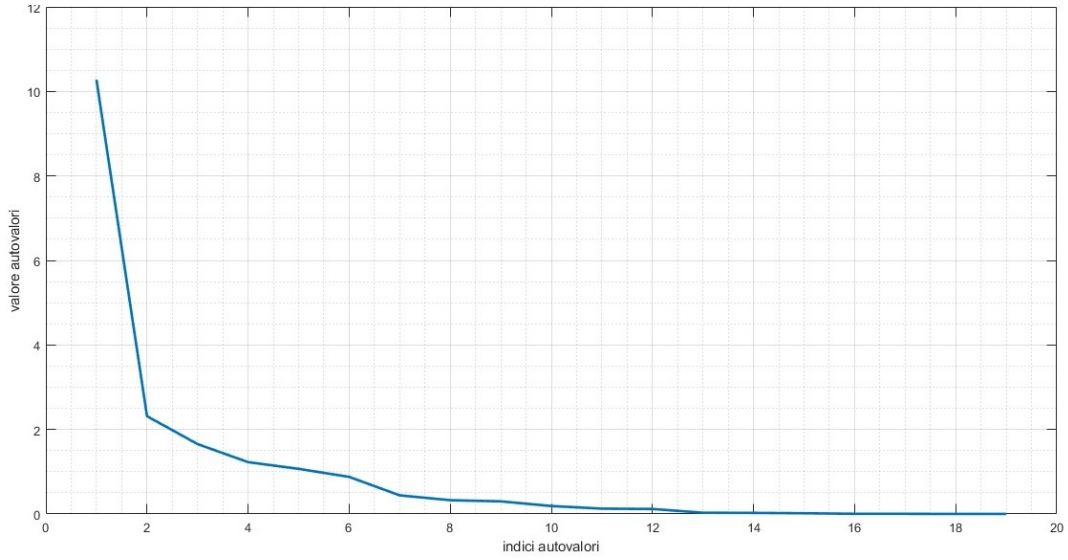


Figura 4.4: Distribuzione degli autovalori ordinati della matrice di correlazione

2. (Percentuale di varianza spiegata)

k	1	**2**	**3**	**4**	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
rv_k	0.5409	**0.6627**	**0.7498**	**0.8144**	0.8706	0.9168	0.9400	0.9571	0.9729	0.9828	0.9894	0.9957	0.9972	0.9986	0.9995	0.9997	0.9999	1.0000	1.0000

3. (Autovalori maggiori della media)

k	**1**	**2**	**3**	**4**	**5**	6	7	8	9	10	11	12	13	14	15	16	17	18	19
λ_k	**10.2768**	**2.3146**	**1.6544**	**1.2269**	**1.0680**	0.8778	0.4408	0.3263	0.2986	0.1883	0.1266	0.1182	0.0298	0.0257	0.0172	0.0046	0.0035	0.0012	0.0005

L'analisi congiunta di questi 3 metodi, permette di prendere anche solo le prime 3 componenti, rappresentando in questo modo il 75% della varianza.
Si guardano ora le componenti principali (colonne di V) per capire quali variabili rappresentano maggiormente, ovvero si cercano i valori più alti in modulo di ciascun autovettore v_i.

	v_1	v_2	v_3	v_4	v_5	v_6
sesso	**-0.2743**	0.0426	0.0836	-0.2104	0.2044	-0.1366
età	0.1573	0.1430	-0.0435	-0.1368	0.5426	0.5398
altezza	**0.2777**	0.1634	-0.1188	-0.1132	-0.0954	0.1491
peso	**0.2973**	0.0846	-0.0021	0.0561	0.1484	0.0503
altezza stream	**0.2774**	0.1689	-0.1094	-0.0791	-0.1341	0.1576
dist.biac.	0.2651	0.1078	0.0067	0.0027	-0.0919	0.2744
larg.spal.stream	0.2426	0.0764	0.2110	0.1097	0.1855	0.0558
BMI	0.2211	-0.0392	0.1284	0.2383	0.4746	-0.1133
PB 50	**-0.2775**	0.1006	0.0171	0.3063	-0.0256	0.1229
FINApoints	0.0979	-0.1125	0.0940	-0.7334	0.2642	-0.3607
Ft	**0.2903**	-0.1337	0.1158	0.0328	-0.1693	-0.1296
Vmax	0.2716	-0.1293	-0.1484	-0.1291	-0.2692	0.0418
Fp	**0.2913**	-0.1592	0.0948	0.0713	-0.0992	-0.1040
Ka	0.2222	-0.0792	**0.4020**	0.2666	0.0693	-0.2622
Kp	0.2166	0.0868	**-0.3108**	0.2750	0.1394	-0.4618
Kcin	-0.0322	-0.1973	**0.7040**	-0.0347	-0.1365	0.2168
SR vmax	0.0820	**-0.5542**	-0.2777	0.0601	0.0693	0.1206
SL vmax	0.1999	**0.3728**	0.0928	-0.1889	-0.3343	-0.0744
SR Ft	0.0784	**-0.5611**	-0.1039	-0.0645	-0.0754	0.1597

Figura 4.5: Rappresentazione delle prime sei colonne di V

Dalla Figura 4.5, si nota che la prima componente principale è associata soprattutto a 'sesso', 'altezza', 'peso', 'altezza in stream', 'PB sui 50', 'Ft' e 'Fp', rappresentando così tutte le "dimensioni": categoriale, antropomentrica e prestativa; la seconda componente principale è associata soprattutto a 'SR vmax', 'SL vmax' e 'SR Ft', rappresentando così la "dimensione" prestativa relativa alla

bracciata; la terza componente principale è associata soprattutto a 'Ka', 'Kp' e 'Kcin', rappresentando così la "dimensione" prestativa relativa alla resistenza idrodinamica.
Quest'ultima osservazione, si può dedurre anche analizzando alcuni grafici di disperione tra le prime tre componenti principali e le variabili originarie standardizzate.

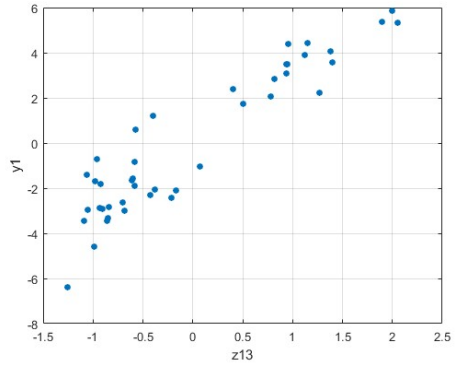


Figura 4.6: Relazione tra la prima componente principale e 'Fp'

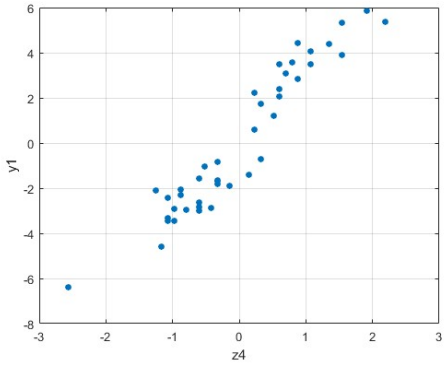


Figura 4.8: Relazione tra la prima componente principale e 'peso'

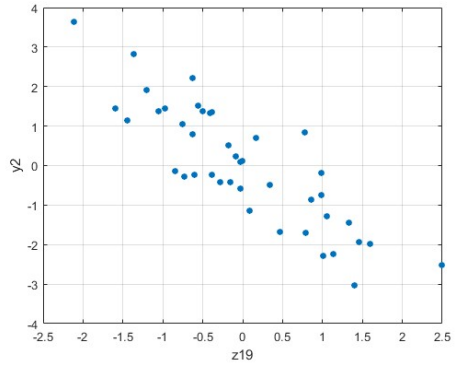


Figura 4.7: Relazione tra la seconda componente principale e 'SR Ft'

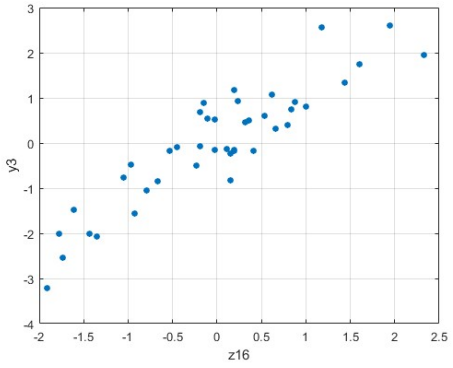


Figura 4.9: Relazione tra la terza componente principale e 'Kcin'

In Figura 4.6, in Figura 4.8 e in Figura 4.9 l'andamento dei punti è approssimabile a una retta crescente, a conferma del fatto che la prima componente principale rappresenta bene la 'forza propulsiva' ($v_{13,1} = 0.2913$) e il 'peso' ($v_{4,1} = 0.2973$) e che la terza componente principale rappresenta bene il 'drag cinetico' ($v_{16,3} = 0.7040$).
In Figura 4.7, invece, l'andamento dei punti è approssimabile a una retta decrescente, a conferma del fatto che la seconda componente principale rappresenta bene la 'frequenza di bracciata al full tethered', ma è inversamente proporzionale a essa ($v_{19,2} = -0.5611$).

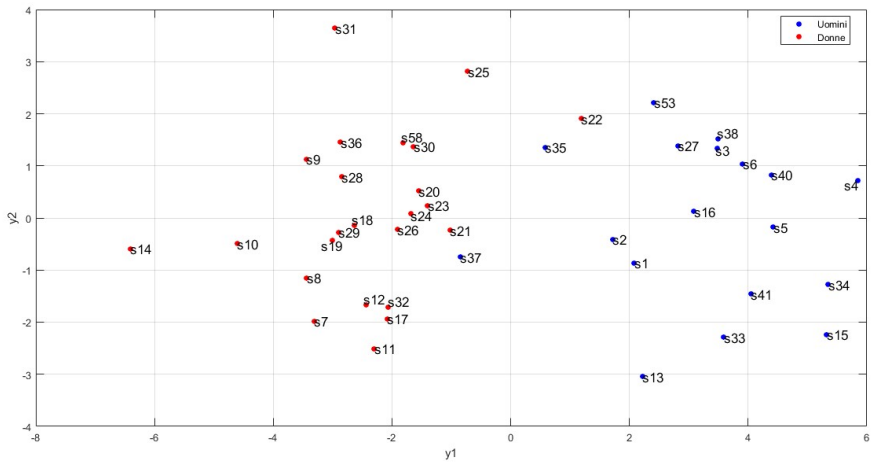


Figura 4.10: Grafico delle prime due componenti principali

In Figura 4.10 è mostrato il grafico 2D dei dati. Più precisamente sulle ascisse è rappresentata la prima componente principale y_1 e sulle ordinate la seconda y_2, le quali complessivamente spiegano il 66.27% della varianza totale.
Si nota come la distribuzione dei punti sia abbastanza diffusa e priva di raggruppamenti netti, nonostante siano abbastanza distinguibili il gruppo delle donne da quello degli uomini, evidenziati con colori differenti. Ciò indica che y_1, y_2 riescono a rappresentare in maniera abbastanza efficiente l'informazione del dataset, ma che le differenze tra i soggetti si distribuiscono in modo continuo senza formare sottogruppi omogenei o cluster separati.

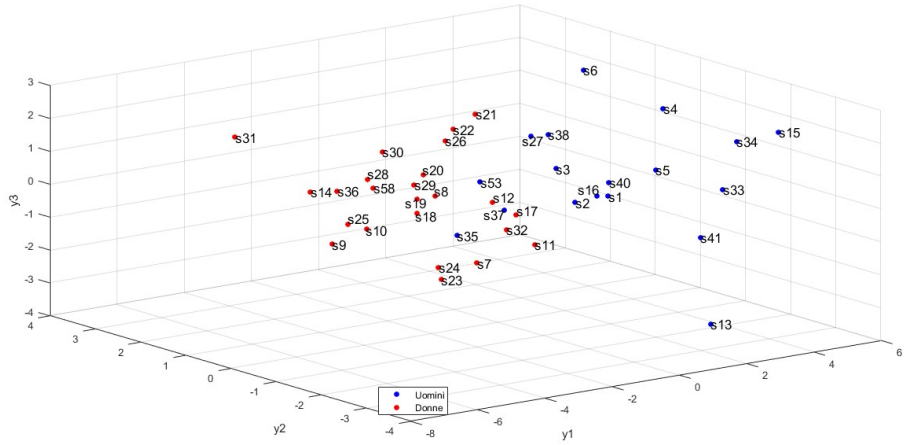


Figura 4.11: Grafico delle prime tre componenti principali

In Figura 4.11 è mostrato il grafico 3D dei dati. Più precisamente sugli assi vengono rappresentate le prime tre componenti principali y_1, y_2, y_3, le quali complessivamente spiegano il 74.98% della varianza totale. L'aggiunta di y_3 consente di migliorare la separabilità visiva di alcune osservazioni e cogliere ulteriori sfumature di variabilità legate a caratteristiche secondarie. Tuttavia la distribuzione continua a non mostrare una netta suddivisione in gruppi distinti nonostante siano abbastanza distinguibili il gruppo delle donne da quello degli uomini, evidenziati con colori differenti.

Quest'analisi grafica suggerisce che la variabilità in questo dataset è progressiva, cioè le differenze tra i soggetti considerati derivano da combinazioni di più caratteristiche che si intrecciano lungo diverse "direzioni" della varianza.

4.2.4 Complete linkage

Ai fini dell'analisi con Complete Linkage, introdotto nella Sezione 3.2, sono state utilizzate quattro distanze differenti quali Euclidea, Cityblock, Cosine e Mahalanobis (viste nella Definizione 3.1.5), in quanto la forma e la scala del dendrogramma dipendono fortemente dal tipo di metrica adottata.

1. *(EUCLIDEA)* Questa distanza è in generale la scelta più solida e accurata su dati standardizzati.

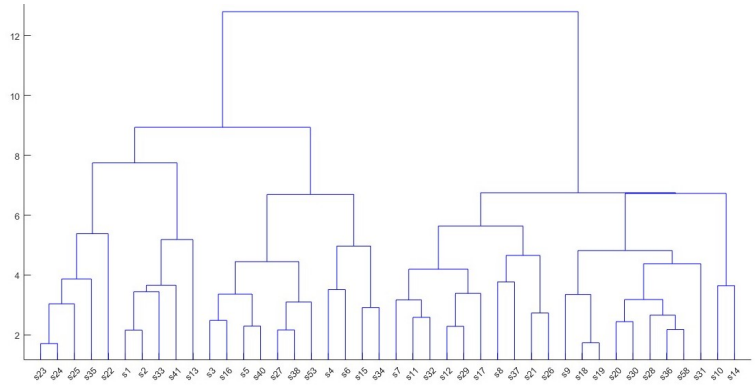


Figura 4.12: Dendrogramma per osservazioni con distanza euclidea

In Figura 4.12 si nota che la struttura del dendrogramma è ben bilanciata: le altezze di fusione aumentano gradualmente, segno che la similarità tra soggetti è lineare. I gruppi risultano compatti e interpretabili. Si distinguono 3–4 cluster principali, separati a livelli di distanza attorno a 8–10, che si articolano ulteriormente in 7-8 sottogruppi, separati attorno a 5–6.

2. *(CITYBLOCK)* Questa distanza è meno sensibile a outlier o a singole variabili con varianza elevata, in quanto pone uguale enfasi su tutte le coordinate.

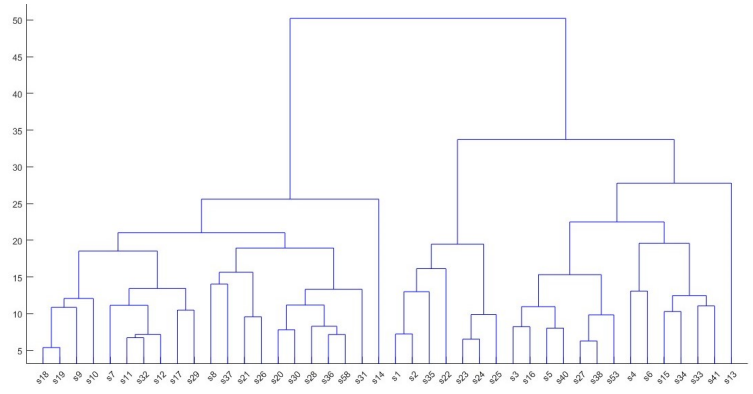


Figura 4.13: Dendrogramma per osservazioni con distanza cityblock

In Figura 4.13 si nota che la struttura del dendrogramma è simile a quella euclidea: le altezze di fusione sono più omogenee, segno che l'uso di questa distanza ha ridotto l'effetto di soggetti "estremi". I gruppi risultano leggermente più equilibrati. Si distinguono 3–4 cluster principali, separati a livelli di distanza attorno a 25-30, che si articolano ulteriormente in 7 sottogruppi, attorno a 20.

3. *(COSINE)* Questa distanza valuta la direzione piuttosto che la magnitudine (due soggetti con le stesse proporzioni risulteranno vicini).

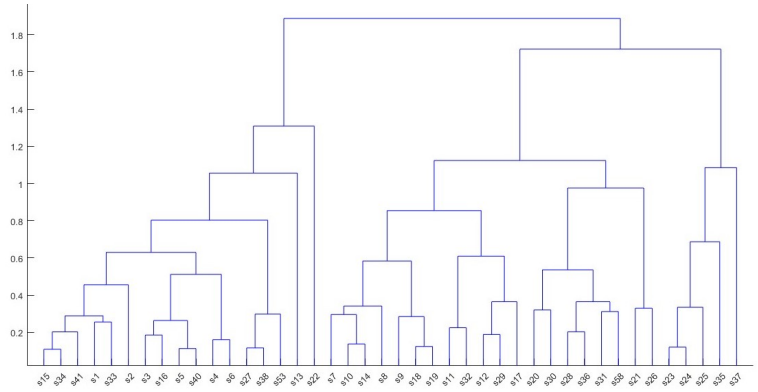


Figura 4.14: Dendrogramma per osservazioni con distanza cosine

In Figura 4.14 si nota che la struttura del dendrogramma è molto diversa da quella euclidea: le altezze di fusione sono molto più basse, segno che questa distanza valuta l'angolo tra i profili e non le differenze assolute. I piccoli gruppi sono numerosi e si uniscono tardi e alcuni gruppi sono separati o ricombinati rispetto ai precedenti. Si distinguono 3–4 cluster principali, separati a livelli di distanza attorno a 1.2-1.4, che si articolano ulteriormente in 10 sottogruppi attorno a 0.8.

4. *(MAHALANOBIS)* Questa distanza è teoricamente più raffinata dell'euclidea e in grado di eliminare la ridondanza tra variabili correlate, ma instabile nel momento in cui ci sono variabili fortemente correlate (come nel nostro caso).

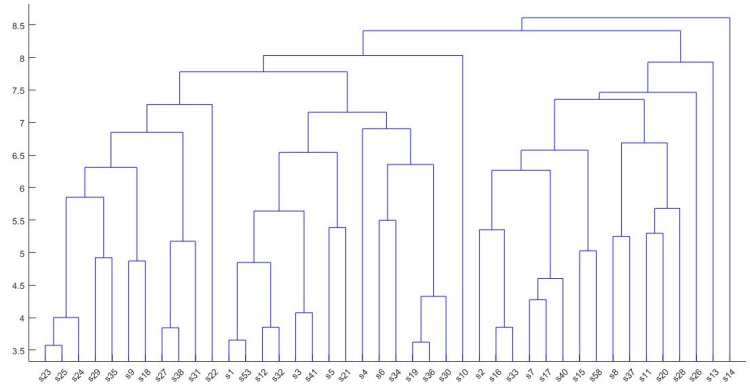


Figura 4.15: Dendrogramma per osservazioni con distanza Mahalanobis

In Figura 4.15 si nota che la struttura del dendrogramma è irregolare e schiacciata: le altezze di fusione sono molto simili tra loro senza che ci siano chiari salti, segno che la matrice di covarianza è molto correlata. Si distinguono 6-7 cluster principali, separati a livelli di distanza attorno a 7.5, che si articolano ulteriormente in 28 sottogruppi attorno a 5.

4.2.5 K-medie

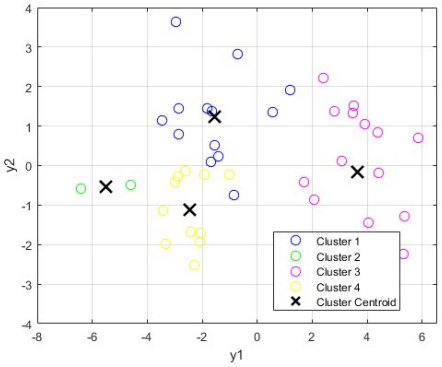
Ai fini dell'analisi con K-medie, introdotto nella Sezione 3.3, come nel Complete Linkage, sono importanti le scelte di inizializzazione dell'algoritmo.
Per individuare il numero di 'Replicates'=riavvii da fare per garantire una maggior precisione del risultato, si stampa in Matlab una tabella Replicates-SSE più bassa e si cerca di capire quanto cambia quest'ultima se si aumenta il numero di riavvii. Si ottiene che il numero ottimale di riavii è 20 e si fissa.
Altro parametro da inizializzare è il numero di gruppi da ottenere che, coerentemente con i risultati ottenuti dal Complete Linkage, viene fissato a 4 e 8.
Per quanto riguarda la distanza per identificare i gruppi, sono state considerate SqEuclidean, Cityblock e Cosine.

Definizione 4.2.1. La distanza SqEuclidean è definita come:

$$d(x, y) = (x - y)^T (x - y) \quad \forall x, y \in \mathbb{R}^p.$$

1. *SQEUCLIDEAN - 4 CLUSTER*

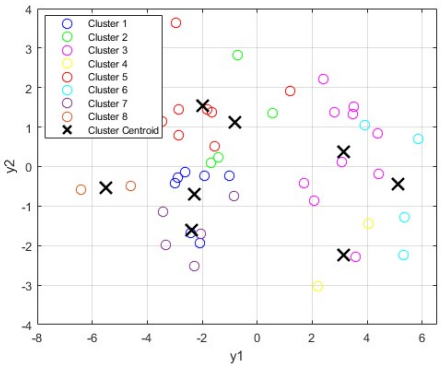
group1	s7,s8,s9,s10,s11, s12,s14,s17,s18,s19, s21,s26,s29,s32,s37
group2	s4,s6,s15,s34
group3	s1,s2,s3,s5,s13,s16,s27, s33,s35,s38,s40,s41,s53
group4	s20,s22,s23,s24,s25, s28,s30,s31,s36,s58



I cluster sono ben compatti e separati e le frontiere appaiono regolari. Questa risulta essere la soluzione più stabile e interpretabile e coincide in gran parte con i gruppi del Complete Linkage euclideo.

2. *SQEUCLIDEAN - 8 CLUSTER*

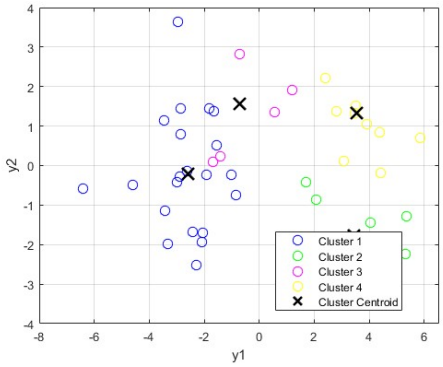
group1	s21,s26
group2	s7,s11,s12,s17,s32
group3	s22,s27,s38,s40,s53
group4	s10,s14
group5	s4,s6,s15,s34
group6	s1,s2,s3,s5, s13,s16,s33,s41
group7	s8,s9,s18,s19,s20,s28, s29,s30,s31,s36,s37,s58
group8	s23,s24,s25,s35



L'aumento da 4 a 8 gruppi porta a una suddivisione coerente dei cluster originari: i grandi gruppi si spezzano in sottogruppi interni più piccoli ma compatti, ben distinti lungo la prima e la seconda componente principale. La distribuzione dei centroidi è equilibrata, segno che la partizione è numericamente stabile.

3. *CITYBLOCK - 4 CLUSTER*

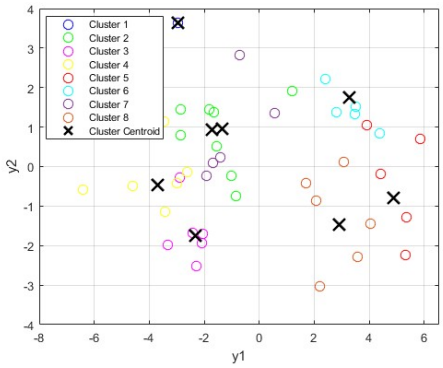
group1	s7,s8,s9,s10,s11, s12,s14,s17,s18,s19, s20,s21,s26,s28,s29, s30,s31,s32,s36,s37,s58
group2	s1,s2,s13,s15, s33,s34,s41
group3	s22,s23,s24,s25,s35
group4	s3,s4,s5,s6,s16, s27,s38,s40,s53



La struttura è simile all'euclidea ma con cluster leggermente più equilibrati (alcune osservazioni cambiano gruppo). Le frontiere sono più "rettangolari" (spezzate e più ortogonali rispetto agli assi) e le fusioni tra cluster centrali avvengono a distanze minori. Essendo questa distanza meno sensibile ai valori anomali, considera le differenze coordinate senza amplificare i picchi.

4. *CITYBLOCK - 8 CLUSTER*

group1	s31
group2	s20,s21,s22,s28, s30,s36,s37,s58
group3	s7,s11,s12,s17,s29,s32
group4	s8,s9,s10,s14,s18,s19
group5	s4,s5,s6,s15,s34
group6	s3,s27,s38,s40,s53
group7	s23,s24,s25,s26,s35
group8	s1,s2,s13,s16,s33,s41

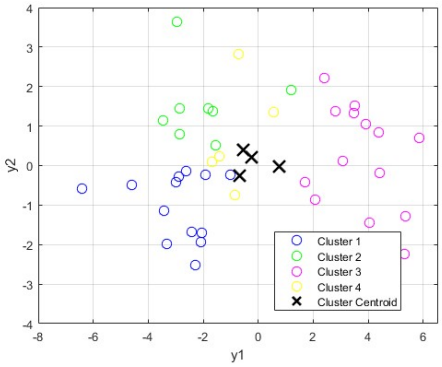


L'aumento da 4 a 8 gruppi porta a una struttura più granulare, suddividendo i cluster principali in sottogruppi più piccoli e compatti. Questa scelta non modifica sostanzialmente la struttura generale dei dati, ma si

notano alcuni confini meno netti, dovuti al fatto che alcuni centroidi sono più vicini tra loro. Inoltre, il numero di membri per gruppo risulta essere più uniforme ed emergono alcuni gruppi con pochissimi elementi.

5. *COSINE - 4 CLUSTER*

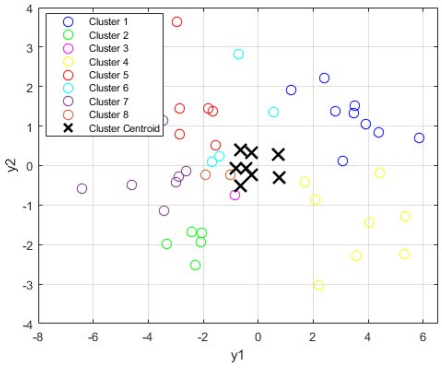
group1	s7,s8,s10,s11,s12, s14,s17,s18,s19, s21,s26,s29,s32
group2	s9,s20,s22,s28, s30,s31,s36,s58
group3	s1,s2,s3,s4,s5,s6, s13,s15,s16,s27,s33, s34,s38,s40,s41,s53
group4	s23,s24,s25,s35,s37



Il grafico mostra cluster più vicini i cui centroidi distano meno rispetto alle precedenti figure: questo perché gli angoli tra vettori sono generalmente piccoli. Alcuni individui che nei precedenti modelli appartenevano a gruppi distinti vengono qui accorpati in maniera differente in quanto ricombinati per similitudine di "pattern".

6. *COSINE - 8 CLUSTER*

group1	s3,s4,s6,s16,s22, s27,s38,s40,s53
group2	s7,s11,s12,s17,s32
group3	s37
group4	s1,s2,s5,s13, ss15,s33,s34,s41
group5	s20,s28,s30,s31,s36,s58
group6	s23,s24,s25,s35
group7	s8,s9,s10,s14,s18,s19,s29
group8	s21,s26



L'aumento da 4 a 8 gruppi porta a una una forte densità centrale: i centroidi sono vicini e i cluster risultano più numerosi e meno separati nel piano, cioè meno chiari visivamente.

Osservazione 4.2.1. Nel complesso, la distanza SqEuclidean rimane la scelta più coerente e stabile per rappresentare la struttura globale del dataset.
La distanza Cityblock conferma la robustezza dei risultati e riduce l'influenza di osservazioni anomale.
La distanza Cosine offre invece una prospettiva complementare, utile per individuare pattern di proporzionalità tra le variabili.

Osservazione 4.2.2. Per eliminare l'influenza dell'unica variabile binaria del dataset (il sesso), l'intera procedura è stata ripetuta dividendo per sesso. Sono stati quindi analizzati separatamente i sottoinsiemi donne e uomini (calcolo correlazioni, analisi delle componenti principali e clustering con Complete Linkage e K-medie).
La varianza totale risulta essere distribuita in un numero maggiore di componenti per le donne e minore per gli uomini. Questo indica che il gruppo femminile presenta una struttura più complessa, dove le caratteristiche antropometriche e prestative non sono dominanti su un'unica direzione di variazione, ma si distribuiscono su più fattori. Ciò è confermato dai grafici 2D delle osservazioni che hanno sugli assi le prime due componenti principali.

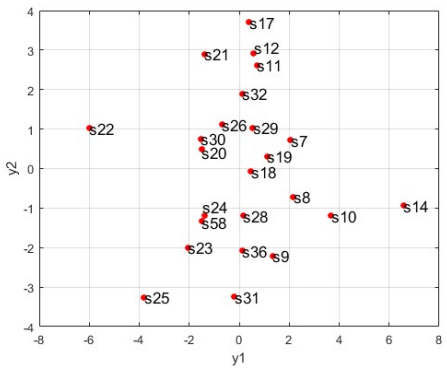


Figura 4.16: Grafico delle prime due componenti principali per donne

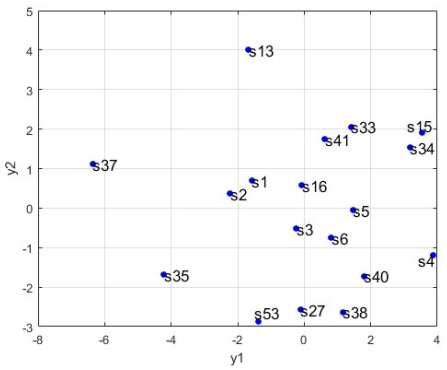


Figura 4.17: Grafico delle prime due componenti principali per uomini

In Figura 4.16 i punti sono dispersi, indicando assenza di sottogruppi ben separabili. La nube è più concentrata e meno varia lungo la prima componente principale, mentre più uniforme lungo la seconda. Le prime due componenti, che sembrano spiegare variabilità differenti, spiegano il 54.03% della varianza.
In Figura 4.17 i punti si dispongono più orientati lungo una retta decrescente. La nube è più larga sulla prima componente principale (che raccoglie una quota maggiore di varianza), mentre più contenuta sulla seconda. Le prime due componenti spiegano il 58.58% della varianza.
Aggiungendo una componente principale sia per le donne che per gli uomini, la varianza spiegata sale rispettivamente a 67.48% e a 74.52%.
Procedendo con l'analisi con Complete Linkage con distanza Euclidea, rivelatasi la più informativa per la tipologia di dati presi in esame, si conferma una struttura più frammentata per le donne e una più coerente per gli uomini, in linea con le differenze di varianza spiegata, evidenziate precedentemente. Si nota che, negli uomini, le aggregazioni avvengono a distanze di fusione maggiori, suggerendo la presenza di cluster più separati.

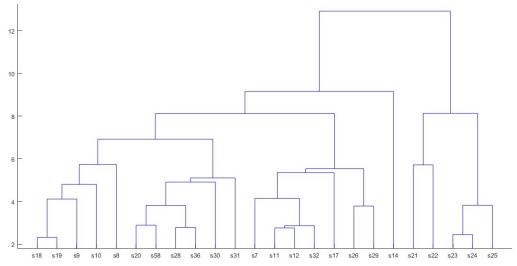


Figura 4.18: Dendrogramma per osservazioni per donne

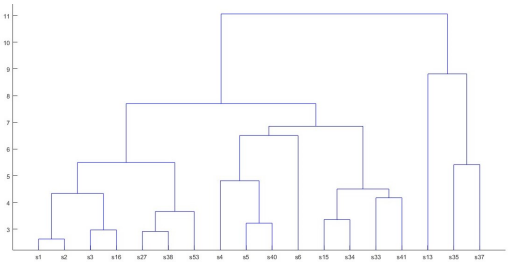


Figura 4.19: Dendrogramma per osservazioni per uomini

Sia in Figura 4.18 che in Figura 4.19 si distinguono 4-5 cluster principali, separati a livelli di distanza attorno a 8.
Si conclude con l'analisi con K-medie con distanza SqEuclidean, mantenendo a 20 il numero di riavvii e fissando, coerentemente con l'osservazione dei dendrogrammi precedenti, a 5 il numero di gruppi. Ancora una volta viene confermata una maggiore eterogeneità interna nel gruppo femminile, intuibile

dalla disposizione meno compatta dei punti e da confini tra gruppi più sfumati in Figura 4.20. Si evidenzia inoltre, in Figura 4.21, che i centroidi risultano essere più ben distinti.

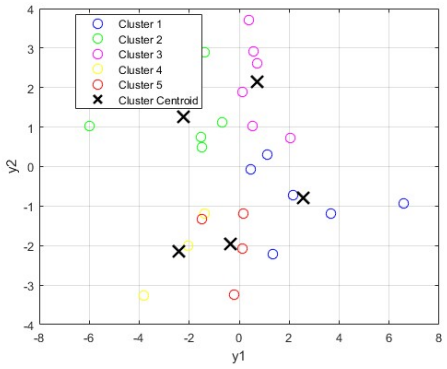


Figura 4.20: Osservazioni delle donne sul grafico delle prime due componenti principali

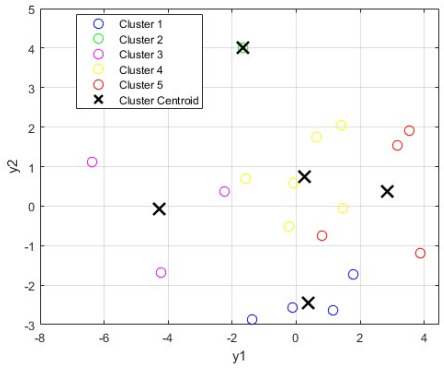


Figura 4.21: Osservazioni degli uomini sul grafico delle prime due componenti principali

Osservazione 4.2.3. Si procede separando le variabili antropometriche da quelle prestative, al fine di verificare se le differenze osservate tra uomini e donne derivino principalmente da fattori antropometrici, prestativi o da una combinazione dei due.
Dall'analisi condotta si evince che, per il gruppo femminile del dataset preso in considerazione, le prestazioni derivano da una combinazione più articolata di fattori morfologici e biomeccanici, mentre negli uomini prevale un modello prestativo più lineare e omogeneo.
La struttura dei dendrogrammi con distanza Euclidea risulta più ramificata e graduale per quanto riguarda le donne e compatta con alcune separazioni più marcate a distanza maggiore, per quanto riguarda gli uomini.
Andando a rifare l'analisi di Complete Linkage con distanza Cosine, si può notare come le donne risultano avere, invece, alcuni pattern e proporzioni simili.
I risultati ottenuti dall'algoritmo di Complete Linkage vegono confermati da quelli ottenuti con le K-medie: per le donne i confini tra i cluster non sono netti con distanza SqEuclidean e, invece, meglio delineati con distanza Cosine. Al

contrario, gli uomini confermano essere distinti in gruppi meglio definiti, con centroidi più separati.

4.3 Analisi con dataset completo

Nonostante i dati disponibili non risultino completi per tutti i soggetti, si può procedere ad un'analisi analoga alla precedente, adottando le dovute accortezze (vedi Sezione 4.1). In questo caso si tengono già divisi i soggetti in base al sesso. Le matrici dei dati saranno quindi: $D \in \mathbb{R}^{27\times47}$ per le donne e $U \in \mathbb{R}^{46\times47}$ per gli uomini.
Siccome il livello di incompletezza delle due matrici risulta essere particolarmente elevato, prima di procedere con l'analisi, è necessario filtrare i dati eliminando righe e colonne con una percentuale di dati mancanti maggiori del 40% della totalità fino ad ottenere due matrici: $D \in \mathbb{R}^{23\times47}$ e $U \in \mathbb{R}^{15\times21}$.

Per entrambe le matrici si sceglie di adottare come metodo di imputazione dei dati la PCA con algoritmo ALS, inizializzando k con un valore appartenente all'intervallo $[2, p-1]$ e uguale circa al 30% di p.
L'algoritmo restituisce:

- $coeff^T_{ALS} \in \mathbb{R}^{p\times k}$ matrice di loadings;
- $score_{ALS} \in \mathbb{R}^{n\times k}$ matrice di scores;
- $mu_{ALS} \in \mathbb{R}^{n\times 1}$ vettore delle medie usate nella PCA.

Questi tre elementi consentono di ricostruire la matrice dei dati imputata:

$$Z_{imp} = score_{ALS} * coeff^T_{ALS} + mu_{ALS}.$$

Una volta ricostruita la matrice dei dati, dopo averla standardizzata, si procede ad un'analisi mediante PCA classica, applicando un procedimento analogo a quello mostrato nello Script 4.2.3.

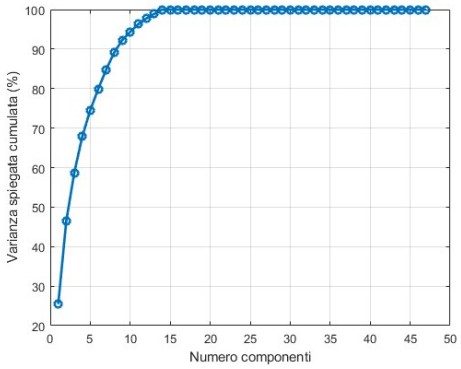


Figura 4.22: Varianza cumulata spiegata per le donne

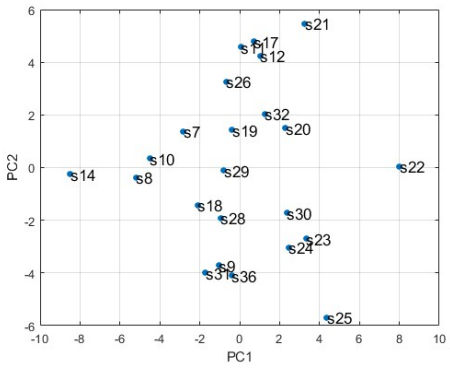


Figura 4.23: Grafico delle prime due componenti principali per le donne

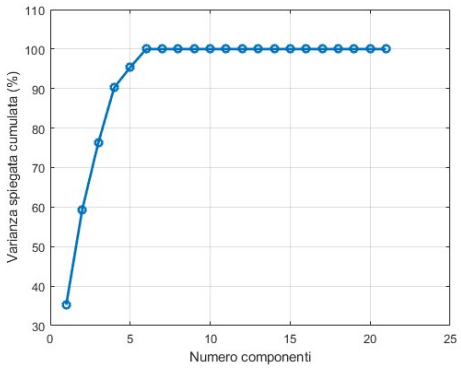


Figura 4.24: Varianza cumulata spiegata per gli uomini

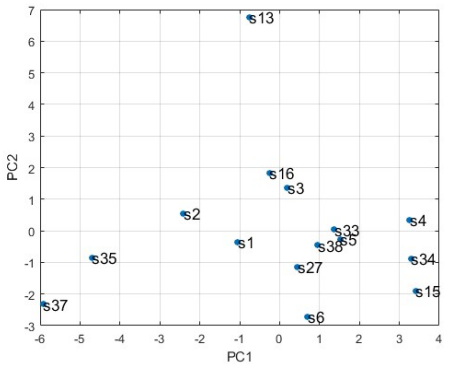


Figura 4.25: Grafico delle prime due componenti principali per gli uomini

Sia in Figura 4.22 che in Figura 4.24 viene evidenziata una crescita molto rapida della varianza cumulata spiegata, al crescere del numero di componenti principali. Nel dataset femminile sono necessarie 5 componenti principali per superare la soglia del 70% di varianza spiegata; mentre in quello maschile lo stesso livello informativo viene raggiunto con sole 3 componenti, indicando una struttura più compatta nello spazio delle variabili. Questa differenza suggerisce che il gruppo maschile presenti pattern più coerenti e correlati tra le variabili, tali da poter essere riassunti in modo più efficiente da un numero ridotto di componenti principali; al contrario del gruppo femminile, che sembra caratte-

rizzato da una maggiore eterogeneità interna.
Osservando i grafici delle prime due componenti principali si evidenziano alcune differenze tra donne e uomini. La Figura 4.23 mostra una buona dispersione dei punti nello spazio bidimensionale, senza sovrapposizioni strette o appiattimenti lungo un asse. Non emergono cluster netti, ma si osservano piccoli sottogruppi con profili simili (ad esempio s11, s12, s17 oppure s9, s31, s36) oltre a soggetti a distanza maggiore dal nucleo centrale del gruppo (ad esempio s14 oppure s22). La Figura 4.25 mostra invece una maggiore dispersione dei punti lungo PC1 rispetto a PC2. Non emergono cluster definiti ma, più che nel caso femminile, si notano soggetti fuori centro, ovvero potenziali outlier.
Ai fini dell'analisi con Complete Linkage, introdotto nella Sezione 3.2, è stata utilizzata solo la distanza Euclidea.

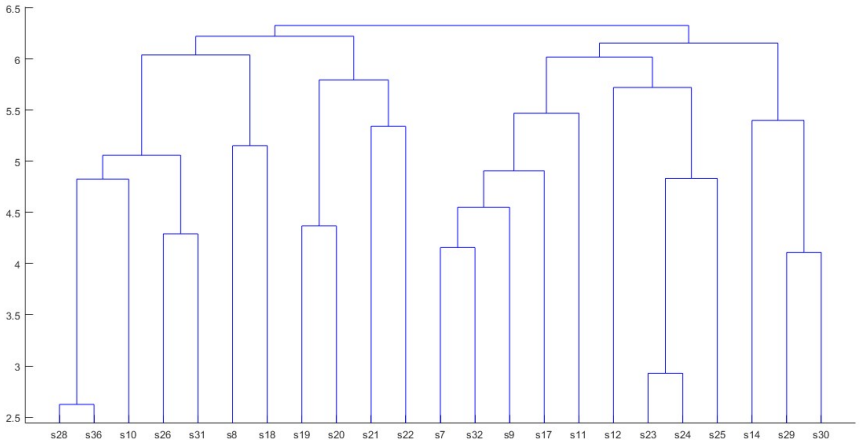


Figura 4.26: Dendrogramma per osservazioni per donne

In Figura 4.26 si osserva una struttura piuttosto articolata, con numerose fusioni che avvengono a quote relativamente alte di distanza. Due coppie di soggetti (s28, s36 e s23, s24) vengono unite molto precocemente rispetto al resto delle osservazioni, suggerendo una forte similarità tra i due soggetti della coppia. Procedendo verso l'alto, le fusioni diventano più eterogenee. Si può notare ad

esempio che s11, s12 e s14 si uniscono solo ad alte distanze, suggerendo che tali soggetti siano più distanti dal resto del gruppo e possano rappresentare valori atipici.

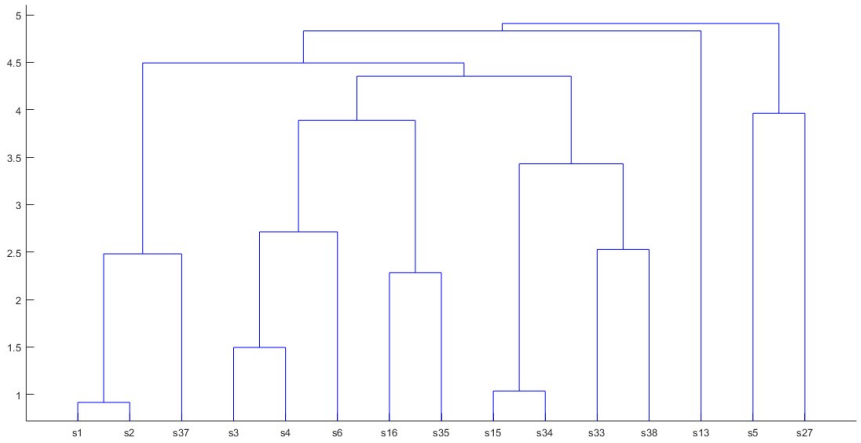


Figura 4.27: Dendrogramma per osservazioni per uomini

In Figura 4.27 si osserva una struttura più compatta e meno frammentata rispetto a quella in Figura 4.26, con fusioni progressive che avvengono a quote più contenute rispetto al caso femminile, che indica una maggiore omogeneità del gruppo. Due coppie di soggetti (s1, s2 e s15, s34) vengono unite prima rispetto al resto delle osservazioni, suggerendo una forte similarità tra i due soggetti della coppia. Solo nelle fasi finali si evidenziano individui più distanti, tra cui s13 e la coppia formata da s5 e s7, suggerendo che tali soggetti possano rappresentare valori atipici.
In sintesi, il Complete Linkage euclideo evidenzia una maggiore eterogeneità nelle donne e una struttura più compatta negli uomini, in piena coerenza con quanto osservato precedentemente.
Per quanto riguarda l'analisi con K-medie, introdotta della Sezione 3.3, si inizializza a 20 il numero di riavvii, a 4 il numero di cluster e come distanza si considera SqEuclidean, coerentemente con l'analisi con Complete Linkage.

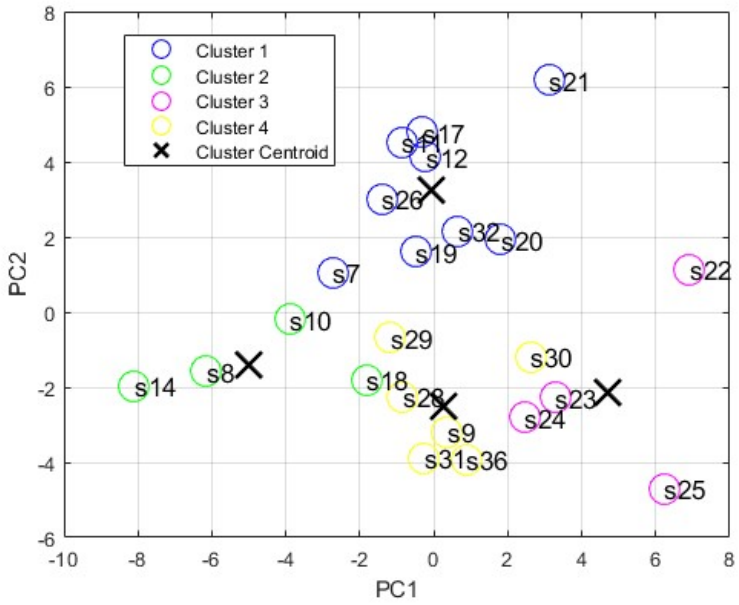


Figura 4.28: Cluster per donne sul grafico delle prime due componenti principali

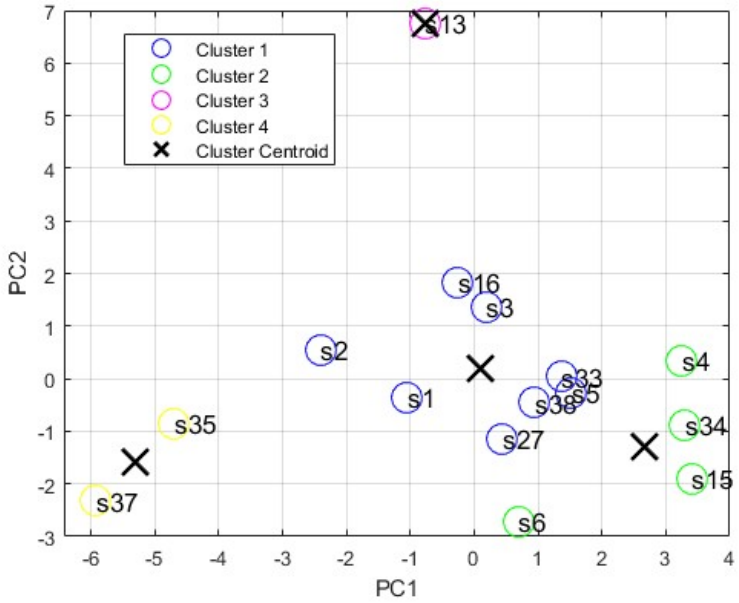


Figura 4.29: Cluster per uomini sul grafico delle prime due componenti principali

In Figura 4.28 si osserva una distribuzione non perfettamente compatta e differenze di densità interna tra i gruppi, coerenti con la maggiore variabilità riscontrata nel dataset femminile. I Cluster 1 e 4 includono un insieme di soggetti omogenei; il Cluster 2 contiene soggetti piuttosto dispersi lungo PC1 e il Cluster 3 contiene casi anomali e distanti rispetto al resto del campione.
In Figura 4.29 si osserva una distribuzione più raccolta lungo PC2, con suddivisione più netta tra i gruppi. I Cluster 1 e 4 includono un insieme di soggetti omogenei; il Cluster 2 contiene soggetti localizzati a PC1 positivi, con s6 spostato rispetto agli altri tre elementi del gruppo e il Cluster 3 contiene solo s13 che, avendo una coordinata in PC2 elevata, risulta essere un potenziale outlier. I risultati ottenuti con K-medie appaiono complementari a quelli derivanti dal Complete Linkage.
In conclusione: le donne presentano una variabilità più ampia, visibile nel grafico di disperione PC1-PC2 (Figura 4.23), nelle fusioni tardive del dendrogramma (Figura 4.26) e nei cluster parzialmente sovrapposti ottenuti da K-medie (Figura 4.28); gli uomini mostrano una struttura più compatta e regolare, con cluster più definiti e stabili in entrambi i metodi, coerenti con una riduzione dimensionale più efficiente (Figura 4.24).

Bibliografia

[1] Adrián Gómez-Sánchez, Raffaele Vitale, Cyril Ruckebusch, and Anna de Juan. Solving the missing value problem in PCA by orthogonalized-alternating least squares (O-ALS). *Chemometrics and Intelligent Laboratory Systems*, 250, 2024.

[2] Alexander Ilin and Tapani Raiko. Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research*, 11, 2010.

[3] Julie Josse and François Husson. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153 no.2, 2012.

[4] Youran Zhou, Sunil Aryal, and Mohamed Reda Bouadjenek. A comprehensive review of handling missing data: Exploring special missing mechanisms, 2024. Preprint, in fase di revisione.